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

Lecture 8: Introduction to Natural Language Processing

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

8.1 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.

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

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

Consider a typical application of *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, an information extraction system will derive important information and insert it into a database, like, given below:

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

8.2 Properties 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 those unique properties so that we are better equipped to understand the processing of natural language(s) through machine.

Following are the unique properties of natural languages [tedbris].

Arbitrariness of the Sign Words relate sounds (or written equivalents) to referents / meanings. There is no systematicity or semantic motivation to this relationship. Onomatopeia is usually a myth (e.g. whisper and French chuchoter are both often said to be onomatopeic), though there are sometimes intuitive commonalities of meaning to words that contain similar sound components. What is common to the meaning of many English words beginning with *gl* (e.g., gland, glucose, glorious, etc.) and can you find some clear exceptions? – look in a dictionary or at text on-line...

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

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 don't have a meaning. The plural morpheme (+s) can be suffixed to three of these words, but is realised as either /s/ and /z/ – so-called allomorphs of the plural morpheme. (Can you explain the exception and the difference?) 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 (*/vlim/, */mbok). Once we allow polysyllabic words (e.g. batter, paddle) is there any restriction on the number of words that can be formed? What is the longest one you know, or can find in a dictionary? What does longest mean in this context?

Syntax Human languages are not just bags of words with no further structure – why not? The organisation of words into sentences is conveyed partly by word structure (endings / inflectional suffixes in English) and arrangement / order. So Kim loves Sandy doesn't mean the same thing as Sandy loves Kim and *loves Kim Sandy doesnt 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 would be like. 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 is shown below:

```
  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 (1a,b), in the following.

1. (a) The cruel boy(s) hit(s) some poor tramp(s) with a sharp rock.
   (b) The cruel, sharp tramp with a 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, imagine trying to discuss politics, philosophy or to explain the design of a computer in such a language!

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. All of the following, sentences are underspecified in this sense. Pronouns, ellipsis (incomplete sentences) and other ambiguities of various kinds all requires additional non-grammatical information to select an appropriate interpretation given the (extra)linguistic context.

1. She smiled
2. I didn't
3. Who?
4. The farmer killed the duckling in the barn
5. Everyone in this room speaks one language
6. Every student thinks he is the cleverest person in JNU
7. Can you open the gate?

Can you contextualise them to give them different meanings and explain how the context resolves the ambiguities?

Whilst the grammatical knowledge required to encode or decode messages in a particular language is circumscribed, the more general inference required to extract messages from utterances is not. Consider the kinds of knowledge you use to make sense of the following dialogues:

1. A: The phone's ringing. B: I'm in the bath.
2. A: John bought a Porsche. B: His wife left him.
3. A: Pint, please. B: Bitter?

You need to know all sorts of culturally specific and quite arbitrary things like the normal location of phones in houses, the semiotics of car brands, and the form of public house transactions, and can make plausible inferences based on these, then these dialogues.

Displacement Most animal 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 learnt (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's very reliable under widely differing conditions, does not require overt tuition, and there isn't 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, '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 6K attested languages we know about, 1K are spoken in Papua New Guinea (an area about the size of Rajasthan). There have probably been 100K- 500K human languages depending on when language first emerged (mostly undocumented, 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. Why?

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, 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 (sociolinguistics).

8.3 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.

8.4 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 8.1.

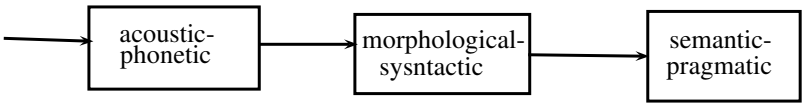

Figure 8.1: The three levels of linguistic analysis.

8.4.1 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 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:* It is useful in understanding the sentence and carry out the conversation. It includes the other persons beliefs and goals.

The figure 8.2 shows the stages of analysis in processing Natural language.

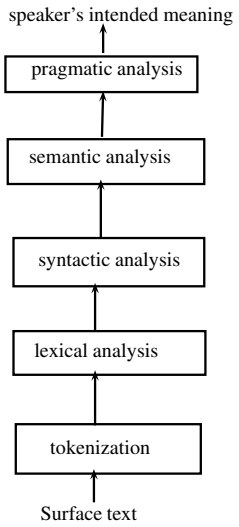

Figure 8.2: Stages in Natural Language Processing.

The concepts of phonological and morphological we have already discussed in details in the previous chapter. Hence, we will be more concerned in our discussions about the remaining domains, i.e., syntax, semantics, pragmatics, etc.

8.4.2 Syntax

Syntax concerns the way in which words can be combined together to form (grammatical) sentences; e.g., revolutionary new ideas appear infrequently is grammatical in English, colourless green ideas sleep furiously is grammatical but nonsensical, whilst **ideas green furiously colourless sleep* is ungrammatical too. (Linguists use asterisks to indicate ungrammaticality, or illegality given the rules of a language.) Words combine syntactically in certain orders in a way which mirrors the meaning conveyed; eg. 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))

8.4.3 Semantics

Semantics is about the manner in which lexical meaning is combined morpho- logically 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 capitalised 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.

8.4.4 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.*

General Knowledge also plays an important role in language interpretation; for example, if I say Lovely day to you whilst we are both being soaked by heavy rain, you will use knowledge that people do not usually like rain to infer that I am being ironic. Similarly, the referents of names and definite descriptions, if not determined situationally, are determined through general knowledge which may be widely shared or not; e.g., the prime minister, Bill, my friend with red hair. Pronoun reference can also often only be determined using general knowledge; e.g., Kim looked at the cat on the table. *It was furry / white / varnished / fat / china / frisky . .*

8.4.5 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 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: 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.

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.

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.

Nonlinguistic 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.

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.

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 F0 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.

8.5 English Language Morphology

It is very useful to be able to analyse words into morphemes and determine their constituents, e.g., their part-of-speech. What follows is a brief outline of how to do this.

Affixation Affixes can be added to word stems (lemmas or headwords with some abstraction to account for spelling / sound change modifications). Combining free and bound (allomorphs of) morphemes (stems and affixes) usually involves spelling changes – *able* → *ability*, *change* → *changing*.

Inflectional suffixes like *+s, +ed* or *+ing* create variants of the same part-of-speech as the stem / headword, e.g. *boy+s N-sg | pl*, *think+s* V-not3sg— 3sg, *think+ing* V-bse— prog, etc. The change in meaning associated with inflectional suffixes relates to the syntactic context in which they occur – they affect agreement, tense etc which are properties of sentences and not (just) words. Derivational affixes affect the inherent meaning of words and often change the part-of-speech too, e.g. *teach(er) V | N, old(er)* A | A-comp(arative). There are productive rules about the combination of morphemes to make words and their consequent meaning:

```
((un    ((re program) able)) ity)
((A/A   ((V/V V)          A\ V)) N\ A)

((un   ((re program) able)) ity)
'the-property-of not being-able to-program (x) again'
```

where X/Y means a prefix combines with a Y to make a word of category X and X\ Y is the analogue for suffixes. What is the final category of the word? What is the bracketing indicating? How do the affixes pair up with the meaning elements in the gloss?

These rules can be motivated by distributional analysis using templates like the following:

1. The – +able computer
2. They re+ – the computer
3. The un+ – computer
4. – +ity is not a good feature

English is relatively isolating (not much inflectional morphology), languages like Hungarian, Finnish and Turkish have many variants (often 100s sometimes 1000s) of each verb. Others, like Arabic, use infixation rather than suffixation (or prefixation): ktb, katab etc. - not much in English but *abso+bloody+lutely* etc. However, English has a lot of derivational affixes and many words are morphologically complex. In the limit, the set of words in English is not finitely specifiable because of iterative / recursive derivational affixes, e.g. *great-great-great grandmother, anti-anti-missile, rereprogram*, etc. This also means that in the limit a lexicon cannot be organised like a conventional dictionary but must be more 'active' / generative, integrating (semi-)productive lexical processes.

Another important lexical process is conversion or zero-derivation in which words change class or gain extended meanings by systematic means, e.g. *purchase, cry* V can become nouns denoting the result of the V *act, butter, oil* N can become verbs denoting the act of applying N, and as mentioned above a lot of animal nouns can also denote the edible flesh of the animal – a semi-productive sense extension i.e. conversion process.

Review Questions

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

Exercises

1. What are the similarities and differences between natural human languages and artificial human languages, such as logics or programming languages? Note: use the properties above as a checklist, but also see if you can think of anything I have not mentioned.
2. How the Information Extraction is different from Information Retrieval?
3. Explain the difference between syntax, semantic, and pragmatic analysis.

References

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

[2] Ted Briscoe, *Introduction to Linguistics for Natural Language Processing*, October 6, 2010, Computer laboratory, University of Cambridge, Monograph.