Topic II

Natural Language Processing

What is NLP?

Natural Language Processing (NLP) is a both a modern computational technology and a method of investigating and evaluating claims about human language itself.

Also called **Computational Linguistics** which links to Artificial Intelligence (AI), the general study of cognitive function by computational processes, normally with an emphasis on the role of knowledge representations, that is to say the need for representations of our knowledge of the world in order to understand human language with computers.

What is Text Processing?

Text processing is a process of **manipulating a written text in a way that will be useful** for further processing or higher level of **NLP application**. Text processing might have different scope based on the *application domain* or NLP type.

- *Understanding* text, *identifying* relevant elements, *manipulating* elements of a text and *analyzing* the structure and semantics of text elements is vital.

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy DraG… ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC
Einstein met with	UN	officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Background

Solving the language-related problems, is the main concern of the fields known as Natural Language Processing, Computational Linguistics, and Speech Recognition and Synthesis

Few applications of language processing

- Spelling correction,
- Grammar checking,
- Information retrieval, and
- Machine translation,
- Speech processing, etc.

Knowledge in NLP

- Tasks of being capable of analyzing an incoming audio signal and recovering the exact sequence of words and generating its response require knowledge about **phonetics and phonology**, which can help model how words are pronounced in colloquial speech.
- Producing and recognizing the variations of individual words (e.g., recognizing that *doors* is plural) requires knowledge about **morphology**, which captures information about the shape and behavior of words in context.

Cont..

Syntax: the knowledge needed to order and group words together

I'm I do, sorry that afraid Dave I'm can't.

(Dave, I'm sorry I'm afraid I can't do that.)

Lexical semantics: knowledge of the meanings of the component words

Compositional semantics: knowledge of how these components combine to form larger meanings

data + base = database

Cont..

Pragmatics: the appropriate use of the kind of polite and indirect language.

You're smart! You got 10 out of 100.

Discourse conventions: knowledge of correctly structuring these such conversations (intonation, gesturer, style, speech act, etc)

Dave, I'm sorry I'm afraid I can't do ***that***.

✓The word "that" is referring to something which is not part of the sentences

Knowledge in Language Processing

- ✓**Phonetics and Phonology**: The study of linguistic sounds
- ✓**Morphology**: The study of the meaningful components of words
- ✓**Syntax**: The study of the structural relationships between words
- ✓**Semantics**: The study of meaning
- ✓**Pragmatics**: The study of how language is used to accomplish goals.
- ✓**Discourse**: The study of linguistic units larger than a single utterance.

Methods and Resources

❖Linguistic Knowledge

Linguistic knowledge resources for many languages are utilized: dictionaries, morphological and syntactic grammars, rules for semantic interpretation, pronunciation and intonation.

❖Corpora and Corpus Tools

Large collections of application-specific or generic collections of spoken and written language are exploited for the acquisition and testing of statistical or rule-based language models.

Approach to NLP

❖Rule Based (Hand Crafted Rules)

Develop the rules to process the natural languages based on known facts and exceptions

❖Machine Learning

Capture rules from examples and apply on new instances

- Supervised: learn by comparing with expected output
- Unsupervised: blind learning.

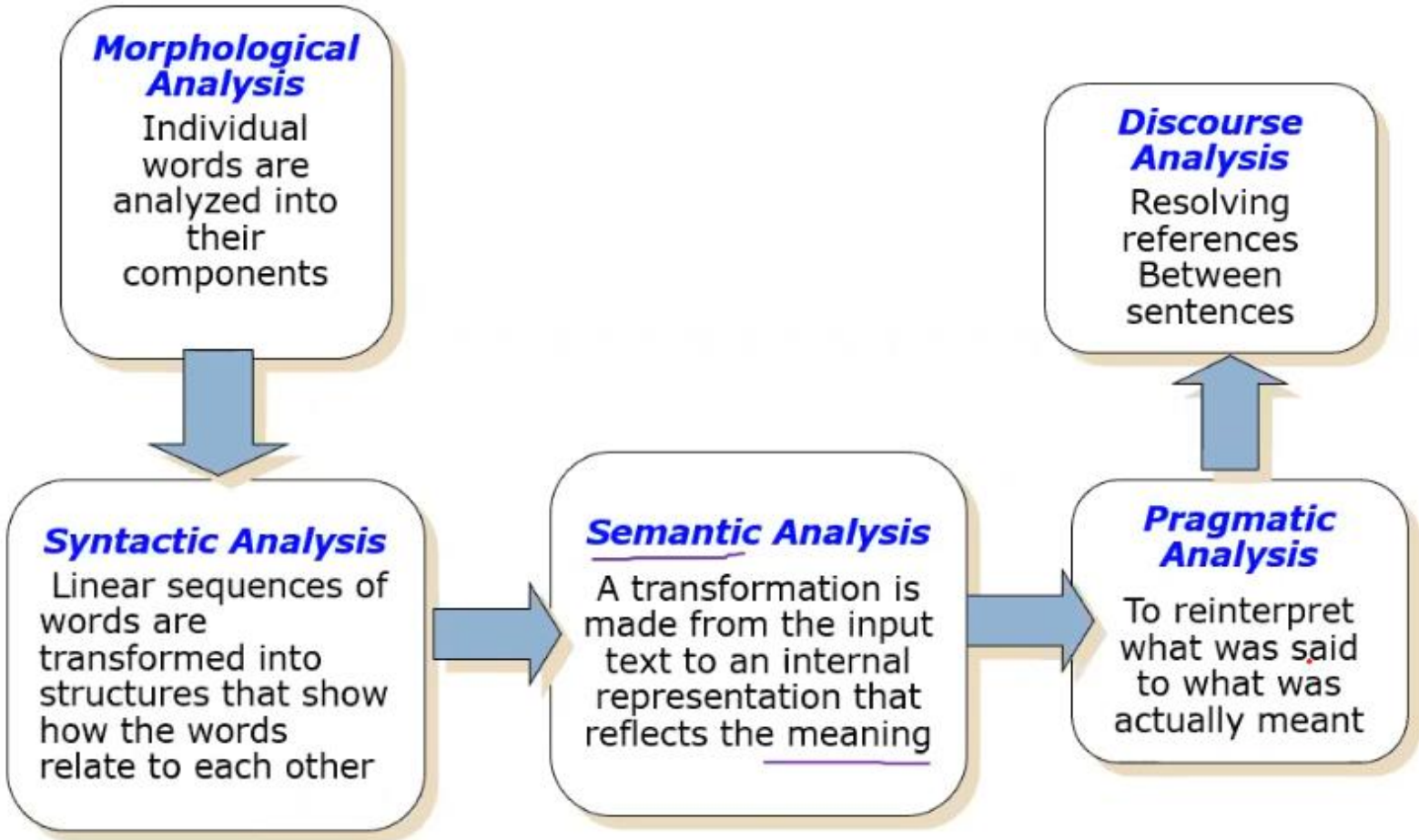
Cont..

- **Machine learning** research has focused on ways to automatically learn the various representations described above;

  - Automata, Rule Systems, Search Heuristics, Classifiers.

These systems can be trained on large corpora and can be used as a powerful modeling technique, especially in places where we don’t yet have good causal models.

Stages of NLP (Textual form)



What is Morphology?

- **Morphology** deals with the syntax of complex words and parts of words, also called ***morphemes***, as well as with the semantics of their lexical meanings.
- Understanding how ***words are formed*** and what ***semantic properties*** they convey through their forms *enables human beings to easily recognize individual words and their meanings* in discourse.

Cont..

Morphology is the branch of linguistics that studies the structure of words.

In English and many other languages, many words can be broken down into parts. For example:

- Un-happi-ness
- Madaxweyn-aha

Cont..

The smallest unit which has a meaning or grammatical function that words can be broken down into are known as **morphemes**.

Morphemes are classified into two types:

- **Free Morphemes**: in Af-Soomaali, words like Madax, Weyne, Inan, Hooyo IWM, and In English words, like; girl, boy, mother, etc. These are words with a complete meaning, so they can stand alone as an *independent word* in a sentence.
- **Bound Morphemes**: These are lexical items incorporated into a word as a *dependent part*. They cannot stand alone, but must be connected to another morpheme to give meaning. For Instance, In Af-Soomali, words like; aha, ooyin, ka, tii IWM. Un & ness.

Word Formation Methods

(1) Affixation is concerned with the way morphemes are connected to existing lexical forms as attachments to show different grammatical feature.

We distinguish affixes of various types:

Prefixes - attached at the beginning of a lexical item or base-morpheme –

e.g. ma, waan, ka, soo

Suffixes – attached at the end of a lexical item

e.g. yaa, sha, ha

(2) Compounding, words can be created by Compounding, which is forming new words from two or more independent words: the words can be free morphemes, words derived by affixation, or even words formed by compounds themselves.

e.g. textbook, database, air-condition

Cont..

(3) Reduplication, which is forming new words either by doubling an entire free morpheme (total reduplication) or part of a morpheme (partial reduplication).

e.g. in *Af-Soomaali*, the word "**jajabay**"

(4) Derivational morphemes create or *derive* new words by changing the meaning or the *word class* of the word (change verb into noun), while

(5) Inflectional morphemes creates a word with similar meaning but more grammatical feature without affecting the word class. For example:

happy → unhappy (Inflectional)

Both words are adjectives, but the meaning changes.

quick → quickness (Derivational)

The affix changes both meaning and word class - adjective to a noun.

Lemmatization and Stemming

- **Lemmatization** usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the **lemma**.
- Lemmatization is the process of identifying lexical/dictionary term after removing all affixes.

Morphological Analysis in its general form involves recovering the LEMMA of a word and all its affixes, together with their grammatical properties.

- **Stemming** a simplified form of morphological analysis – simply find the stem.

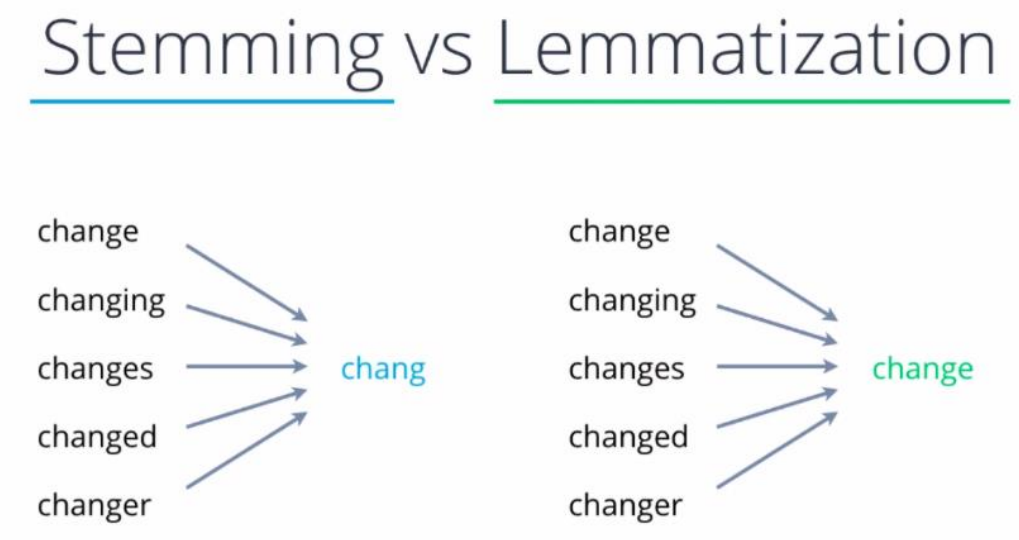
Cont..

- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
- Reduce terms to their stems in information retrieval

Stemming is crude chopping of affixes

- Language dependent

e.g., automate(s), automatic, automation all reduced to automat.



Roots & Stem

Roots:

The root is generally the principle carrier of the lexical meaning of a word, while affixes generally carry grammatical meanings.

For example, in cats, the root cat carries the basic meaning, while -s carries the grammatical information 'plural.'

Stems:

In addition to roots, we also distinguish stems. A stem may be also a root, as cat in cats.

Exercise: Somali Root vs Stem?

Tokenization

Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

The are two types of tokenization:

- Word Tokenization
- Sentence Tokenization

Cont..

they lay back on the San Francisco grass and looked at the stars and their

Type: an element of the vocabulary.

Token: an instance of that type in running text.

How many?

- 15 tokens
- 13 types

Practical Sessions of this Lecture!

First thing first, Install NLTK for your Computer. Then try to do:

Tokenization in NLP

Lemmatization in NLP

Stemming in NLP