Mehran University of Engineering & Technology, Pakistan

NATURAL LANUGAUGE PROCESSING (NLP)

Department of Software Engineering

Introduction

- A subfield of Artificial Intelligence (AI) that deals with the processing of Natural Language.
- The term Natural language refers to languages in common and daily use such as English, French, Spanish, Urdu, Arabic.
- The study of NLP is concerned with developing an intelligent system that can be augmented with capabilities for processing Natural Language & for translating b/w different languages.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

What is NLP?

(From the Natural Language Processing Research Group at the University of Sheffield Department of Computer Science.)

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

- Some prefer the term **Computational Linguistics** in order to capture this latter function, but NLP is a term that links back into the history of 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.

- *Natural Language Processing (NLP) is the use of computers to process written and spoken language for some practical, useful, purpose: to translate languages, to get information from the web on text data banks so as to answer questions, to carry on conversations with machines, so as to get advice about, say, pensions and so on.*

What is NLP?

(From the Natural Language Processing Research Group at the University of Sheffield Department of Computer Science.)

- These are only examples of major types of NLP, and there is also a huge range of lesser but interesting applications, e.g. getting a computer to decide if one newspaper story has been rewritten from another or not.

- NLP is not simply applications but the core technical methods and theories that the major tasks above divide up into, such as Machine Learning techniques, which is automating the construction and adaptation of machine dictionaries, modeling human agents' beliefs and desires etc. This last is closer to Artificial Intelligence, and is an essential component of NLP if computers are to engage in realistic conversations: they must, like us, have an internal model of the humans they converse with."

What is NLP.

(From wikipedia)

- **"Natural language processing** (**NLP**) is a field of computer science concerned with the interactions between computers and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language. Natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate".

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

What is NLP.

(From Sci-Tech Encyclopedia)

- “Computer analysis and generation of natural language text”.
- The goal is to enable natural languages, such as English, French, or Japanese, to serve either as the medium through which users interact with computer systems such as database management systems and expert systems (natural language interaction), or as the object that a system processes into some more useful form such as in automatic text translation or text summarization (natural language text processing).

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

NLP Application Areas

- Machine Translation,.
- Natural Language Text Processing And Summarization.
- User Interfaces.
- Multilingual And Cross Language Information Retrieval (CLIR).
- Speech Recognition.
- Spelling And Grammar Checkers.
- Natural Language Interfaces.
- Spoken Language Control System.
- Automatic Message Understanding And Classification Systems, etc.

The Entire NLP Problem Is Divided In Two Tasks:

- **Processing written text,** using lexical, syntactic, and semantic knowledge of the language as well as the required real world information.
- **Processing spoken language,** using all the information needed above plus additional knowledge about phonology as well as enough added information to handle the further ambiguities that arise in speech.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

SUBFIELDS of NLP

- Natural Language **Generation**
- Natural Language **Understanding**

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

NATURAL LANGUAGE GENERATION

- It is the area of NLP concern with making it easier for you to understand a computer output.
- The ability of computer to generate the output in your natural language allows more understandability of computer's response.
- A natural language program generation program has the following three basic functionalities:
  - The program must decide when to say something
  - The program must decide what to say.
  - The program must decide how to say it.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Natural Language Understanding

- Concerned with the understanding of knowledge specified in Natural Language such that the computer can use it to perform it tasks.
- It is much easier to build a computer system that can do calculus than to build one that can understand language. The reason is that understanding language requires a vast amount of knowledge.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Features of Natural Language That makes it difficult OR useful

- The Problem:
  - English sentences are incomplete descriptions of information that they are intended to convey:

Some Dogs are outside . . . Some dogs are on the lawn. some dogs are on street.	**I called Nancy to ask her to go for shopping.** **She said she'd love to go.** **.** **.** **.** **She was home when I called her.** **She answered the phone.**

- The Good Side:
  - Language allows speakers to be as vague or precise. Leaving things on listener to understand or known already.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

- The Problem:
  - The same expression means different things in different contents.

Where's the water? ….. (in a chemistry lab, it must be pure)

Where's the water? ….. (when you are thirsty, it must be potable)

Where's the water? ….. (dealing with a leaky roof, it can be filthy)

Where's the water? ….. (searching on Indus river from Jamshoro bridge)

- The Good Side:
  - Language lets us communicate about an infinite world using a finite (learnable) number of symbols.

- The Problem:
  - No natural language program can have complete knowledge because of increasing number of new words, expressions are generated quickly.

I'll Fax it to You.

Google this topic at home.

Let's Tweet for today's Session

- The Good Side:
  - Language can evolve as the experiences that we want to talk or communicate about.

- The problem:
  - There are lot of ways to say the same thing.

Mary was born on October 11.

Mary's Birthday is October 11.

- The Good Side:
  - When you know a lot,, facts implies easily to each other. It is intended to be known by user who know a lot.

"I never said she stole my money"

- This demonstrates the importance that stress can play in a sentence.
- Depending on which word the speaker places the stress, this sentence could have several distinct meanings:
- "**I** never said she stole my money" - Someone else said it, but *I* didn't.
- "I **never** said she stole my money" - I simply didn't ever say it.
- "I never **said** she stole my money" - I might have implied it in some way, but I never explicitly said it.
- "I never said **she** stole my money" - I said someone took it; I didn't say it was she.
- "I never said she **stole** my money" - I just said she probably borrowed it.
- "I never said she stole **my** money" - I said she stole someone else's money.
- "I never said she stole my **money**" - I said she stole something, but not my money.

Other Problems in NL Understanding:

- **Ambiguity:**
  - **Lexical Ambiguity**: Multiple word meaning as in following sentences.
    - *The pitcher is angry*
    - *The pitcher is empty*
  - **Syntactic ambiguity**: e.g. *I hit the man with the hammer*
    - How do you interpret it, either I hit the man who was holding the hammer, or I hit the man using a hammer.
  - **Referential Ambiguity**: Unclear antecedents, problem in understanding target of pronouns such as
    - *John hates bill because he sympathized with Sam.*
    - Again who is he, either we mean here John or bill.

Dr. Isma F. Siddiqui (isma.farah@faculty.muet.edu.pk)

- **Imprecision:** Take following sentences
  - ***I waited for taxi for a long time.***
  - ***He died because he didn't eat for a long time.***

➢Here the term **long time** has imprecise meaning, in one sentence it may mean several minutes whereas in the other one it may mean several days.

- **Incompleteness:** Consider the following passage.
  - ***I went to hotel yesterday for dinner. After I paid the bill, I noticed that I had very less money left.***

➢Considering above passage, it is not stated that whether I ate the dinner or not.

- **Inaccuracy:** Inaccuracy can rise due to following factors
  - **Spelling errors**
  - **Incomplete sentences**
  - **Improper punctuation, etc.**

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Solutions:

- Following table shows some possible solutions to above stated problems.

Problem	Solution
Ambiguity	Use of background knowledge and context
Imprecision	Relate idea to a familiar situation
Incompleteness	Complete idea based on expectations
Inaccuracy	Infer based on recognition of familiar pattern

ANALYSIS LEVELS OF NATURAL LANGUAGE UNDERSTANDING

- Natural Language Understanding involves several complex processes. To manage this complexity, linguists have defined different levels of analysis for natural language:
  - **Phonology**
  - **Morphology**
  - **Syntax**
  - **Semantics**
  - **Pragmatics**
  - **Discourse Integration**
  - **Prosody**

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Phonology

- It is used for spoken natural language understanding.
- It is the study of the sounds that makeup words and is used to identify words from sound.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Morphology

- It looks at the ways in which word break down into components and how that affects their grammatical status.
- These include the rules governing the formation of words, such as the effect of prefixes (un-, non-, anti-, etc.) and suffixes (-ing, -Iy, etc.) to modify the meaning of root words.
- For example: the letter “s” on the end of any word can often either indicate plural of it or third person present-tense verb.

Syntax

- It involves applying the rules of the grammar from the language being used.
- It determines the role of each word in a sentence, and enable computers to convert into easily manipulated structures.
- It requires some kind of parsing method.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Semantics

- It involves the examination of the meaning of words and sentences.
- Some methods of semantic analysis makes use of various types of grammars, which are formal systems of rules that attempt to describe the ways that the sentences can be constructed.
- Semantic Analysis must do two important thing:
  - It must map individual words into appropriate objects in the knowledge base or database.
  - It must create the correct structures to correspond to the way the meanings of the individual words.
- Such systems include Conceptual dependency theory and Conceptual Networks.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Pragmatics

- This is the application of human-like understanding to sentences and discourse to determine meanings that are not immediately clear from the semantics.
- It is the study of the ways in which language is used and its effects on the listener
- For example: “Can you tell me the time?”
  - Here “Yes” is not the suitable answer
  - Pragmatics enables a computer system to give a sensible answer to questions like this.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Discourse Integration

- It is study of the structure of extended (multiple sentence) language such as in text or in dialogs.

Prosody

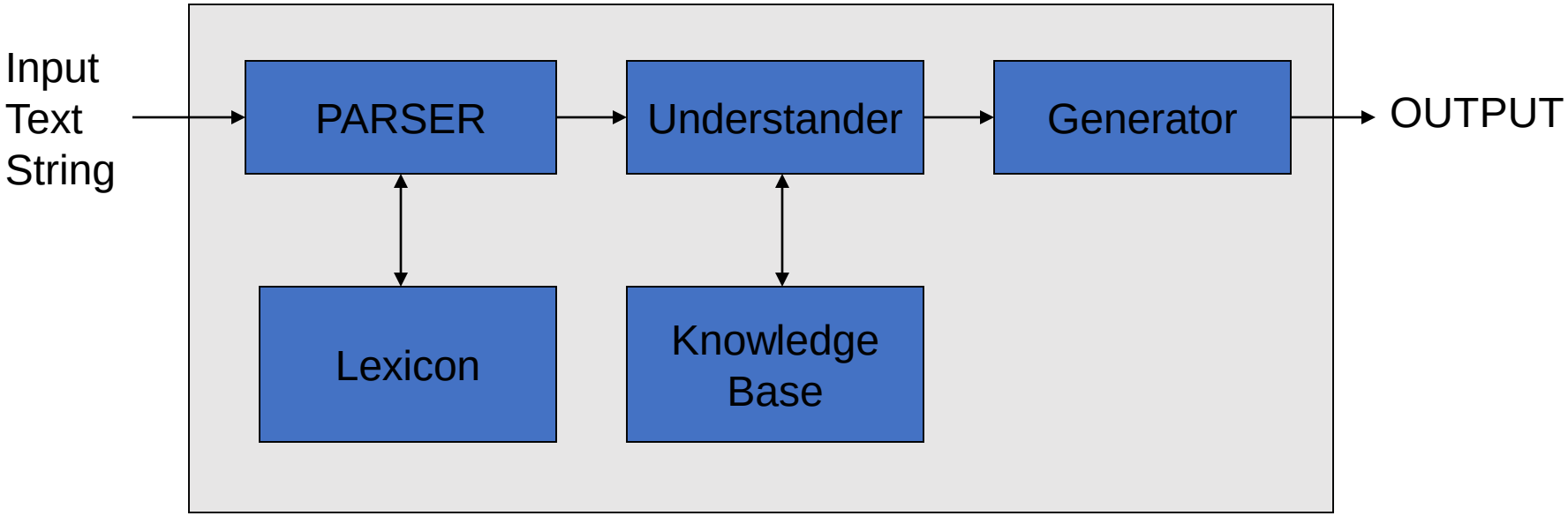
- It deals with the rhythm and intonation of language. Or, The stress and intonation patterns of an utterance.
- This level of analysis is difficult to formalize and often neglected.

Parsing

- Parsing ensures that the computer understands the precise function of each word in a sentence and its relation with other words in the sentence.
- In this process sentence is converted into a hierarchical structure that corresponds to the units of meaning in the sentence.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Natural Language Processing System



Parser

- The parser is a piece of software that analyzes the input sentence syntactically.
- Each word is isolated and its part of speech is identified.
- The parser then maps the words into a structure called the parse tree.
- The parse tree shows the meanings of all of the words and how they are assembled.

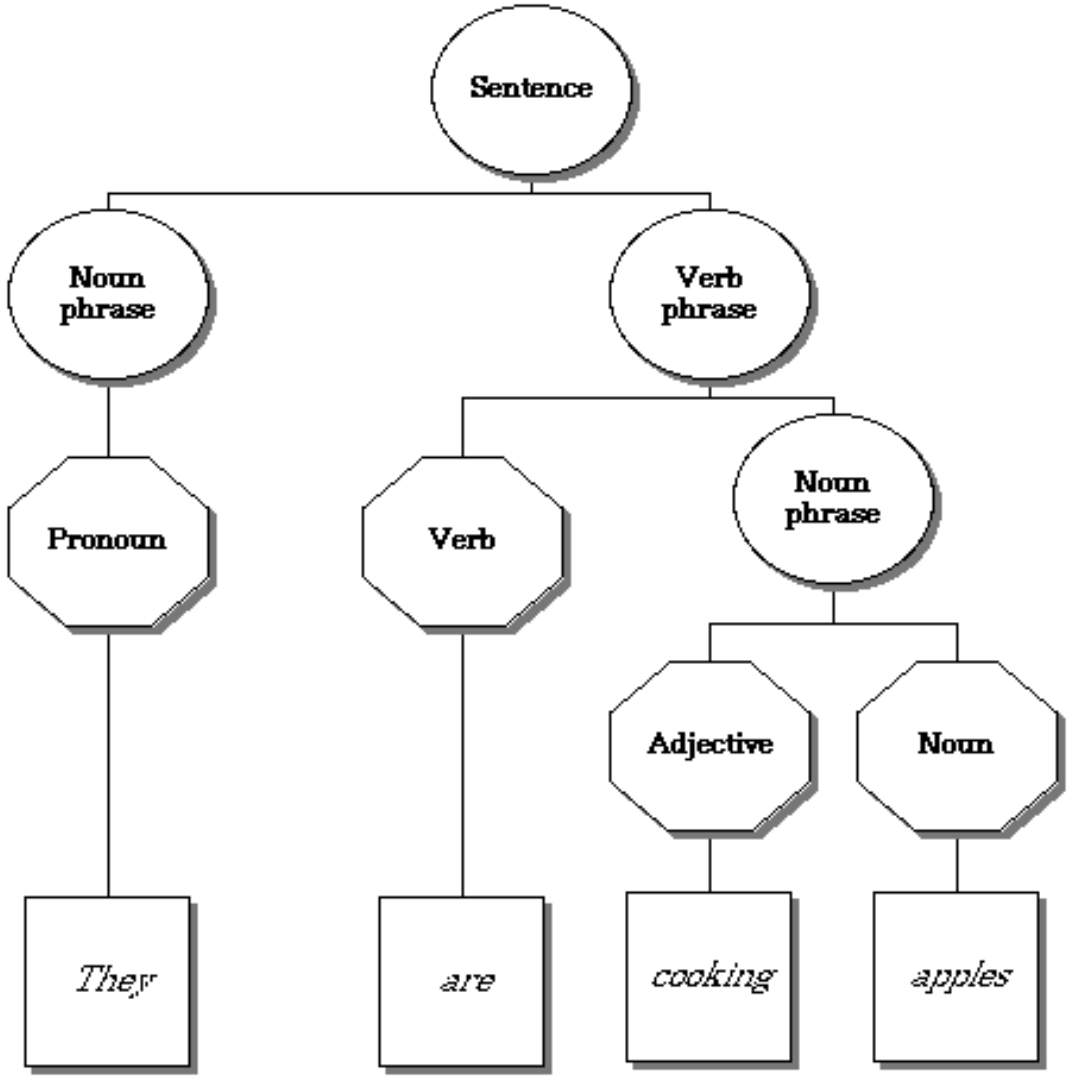
- The job of the parser is to examine each word in a sentence and create the parse tree that identifies all the words and puts them together in the right way.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Working of Parser

- The parser identifies the noun phrase and verb phrase and further breaks them down into other elements.
- This syntactic analysis is first step toward trying to extract meaning from the sentences.
- A sentence is made up of a subject or noun phrase (NP) and a predicate or verb phrase (VP). Shown as:
    - S = NP + VP
- The noun phrase could be single noun, but it usually breaks down further into several additional parts of speech, such as an article (ART) or determiner (D) like *a or this*, and/or an adjective (ADJ) or two, and the main noun.
    - NP = D + ADJ + N
- The noun phrase may even have a prepositional phrase (PP) made up of a preposition (P) such as *of* or *with* and another determiner (D) and noun (N):
    - PP = P + D + N
- The verb phrase (VP) is made up of the verb (V) and often the object of the verb, which is usually another noun (N) and its determiner (D). A prepositional phrase (PP) may also be associated with the verb phrase.
    - VP = V + D + N + PP

"They are cooking apples"



Lexicon

- The lexicon contains all the words that the program is capable of recognizing.
- The parser works closely with the lexicon in doing syntactic analysis.
- The lexicon is composed of a general component as well as a domain or task specific one.
- The lexicon contains the correct spelling of each word and also designates its part of speech.
- For words that can have more than one meaning, the lexicon lists all the various meanings permitted by system.
- The parser and the lexicon work together to pick apart a sentence and then create the parse tree.
- In operation, the parser is largely a pattern matcher. Once the individual word is identified, the parser searches through the lexicon comparing each input word with all of the words stored there. If a match is found, the word is put aside along with the other lexical information., such as part of speech and meaning.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Understander and Knowledge Base

- Semantic analysis is the function of the understander.
- Understander works in conjunction with the knowledge base to determine what the sentence means.
- The knowledge base can be divided conceptually into two parts:
  - The general knowledge base: (general world knowledge and basic linguistic concepts)
  - The domain task-specific knowledge base. (knowledge about entities and computational procedures specific to the application at hand).
- The knowledge base is the primary means of understanding.
- The purpose of understander is to use the parse tree to reference the knowledge base.
- The understander can also draw inferences from the input statements.

Generator

- The generator uses the understood input to create a usable output.
- The understander creates another data structure that represents the meaning and understanding of the sentence and stores it in memory. That data structure can then be used to initiate additional action.
- If NLP is part of a front-end interface, the data structure will be used to create special codes to control another piece of software. It may give software commands to initiate some action.
- In DBMS, the generator would write a program in a query language to begin a search for specific information.
- In other systems, the action may be the generator of an output response such as a sentence or question.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Linguistic tools

- Linguistic analysis of text typically proceeds in a layered fashion.
- Documents are broken up into paragraphs, paragraphs into sentences, and sentences into individual words.
- Words in a sentence are then *tagged* by part of speech and other features, prior to the sentence being *parsed* (subjected to grammatical analysis).
- Thus parsers typically build upon sentence delimiters, tokenizers, stemmers, and part of speech (POS) taggers.
- But not all applications require a full suite of such tools.
- For example, all search engines perform a tokenization step, but not all perform part of speech tagging.

Language Model

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Language Models

- A language model is a probabilistic mechanism for generating text
- Language models estimate the probability distribution of various natural language phenomena
  - sentences, utterances, queries …
- A statistical language model assigns a *probability* to a *sequence of m words* by means of a probability distribution
- Applications in NLP:
  - speech recognition,
  - machine translation,
  - part-of-speech tagging,
  - parsing,
  - information retrieval.
- The goal of Statistical Language Modeling is to build a statistical language model that can estimate the distribution of natural language as accurate as possible.

What is an N-Gram?

- A subsequence of *n items* from a given sequence
  - Unigram: *n*-gram of size 1
  - Bigram: *n*-gram of size 2
  - Trigram: *n*-gram of size 3
- Items:
  - Phonemes
  - Syllables
  - Letters
  - Words
- Number of Items:
  - Unigram, Bigram, Trigram, ...

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Example

the dog smelled like a skunk

- Bigram:
  - "# the", “the dog”, “dog smelled”, “ smelled like”, “like a”, “a skunk”, “skunk#”
- Trigram:
  - "# the dog", "the dog smelled", "dog smelled like", "smelled like a", "like a skunk" and "a skunk #".

N-Gram Model

- A Probabilistic Model for *Predicting* the next Item in such a sequence.
- Why do we want to Predict Words?
  - Chatbots
  - Speech recognition
  - Handwriting recognition/OCR
  - Spelling correction
  - Author attribution
  - Plagiarism detection
  - ...

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

N-Gram Model

- Predicts x_i based on $x_{i-1}, x_{i-2}, \ldots, x_{i-n}$:

$$P(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_{i-n})$$

- NGram *Independence Assumption:*
  - *word is affected only by its "prior local context" (last few words)*
  - *Advantages:*
    - *Massively simplifies the problem of learning the language model*
    - *because of the open nature of language, it is common to group words unknown to the language model together*

Morphology

- Morphology is the study of the internal structure of words, of the way words are built up from smaller meaning units.
- Morpheme:
  - The smallest meaningful unit in the grammar of a language.
- Two classes of morphemes
  - Stems: “main” morpheme of the word, supplying the main meaning (i.e. *establish* in the example below)
  - Affixes: add additional meaning
    - Prefixes: Antidisestablishmentarianism
    - Suffixes: Antidisestablishmentarianism
    - Infixes: hingi (*borrow*) – humingi (*borrower*) in Tagalog
    - Circumfixes: sagen (*say*) – gesagt (*said*) in German

Dr. Isma F. Siddiqui (isma.farah@faculty.muet.edu.pk)

Morphology: examples

- **Unladylike**
  - The word *unladylike* consists of three morphemes and four syllables.
  - Morpheme breaks:
    - un- 'not'
    - lady '(well behaved) female adult human'
    - -like 'having the characteristics of'
  - None of these morphemes can be broken up any more without losing all sense of meaning. *Lady* cannot be broken up into "la" and "dy," even though "la" and "dy" are separate syllables. Note that each syllable has no meaning on its own.
- **Dogs**
  - The word *dogs* consists of two morphemes and one syllable:
    - dog, and
    - -s, a plural marker on nouns
  - Note that a morpheme like "-s" can just be a single phoneme and does not have to be a whole syllable.
- **Technique**
  - The word *technique* consists of only one morpheme having two syllables.
  - Even though the word has two syllables, it is a single morpheme because it cannot be broken down into smaller meaningful parts.

Adapted from http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsAMorpheme.htm

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Types of morphological processes

- ***Inflection***:
  - Systematic modification of a root form by means of prefixes and suffixes to indicate grammatical distinctions like singular and plural.
  - Stems: also called lemma, base form, root, lexeme
  - Doesn't change the word class
  - New grammatical role
  - Usually produces a predictable, non idiosyncratic change of meaning.
    - run → runs | running | ran
    - hope+ing → hoping hop → hopping

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Types of morphological processes

- ***Derivation***:
  - Ex: *compute* → *computer* → *computerization*
  - Less systematic that inflection
  - It can involve a change of meaning
    - *Wide* → *Widely*
    - Suffix *en* transforms adjective into verbs
      - *Weak* → *weaken*, *soft* → *soften*
    - Suffix *able* transforms verbs into adjective
      - *Understand* → *Understandable*
    - Suffix *er* transforms verbs into nouns (nominalization)
      - *teach* → *teacher*
  - Difficult cases:
    - *building* → from which sense of "*build*"?

Dr. Isma F. Siddiqui (isma.farah@faculty.muet.edu.pk)

Types of morphological processes

- *Compounding*:
  - Merging of two or more words into a new word
    - *Downmarket, (to) overtake*

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Stemming

- The removal of the inflectional ending from words (Cut off any affixes)
  - *Laughing*, *laugh*, *laughs*, *laughed* → *laugh*
  - Problems
    - Can conflate semantically different words
      - *Gallery* and *gall* may both be stemmed to *gall*
  - A further step is to make sure that the resulting form is a known word in a dictionary, a task known as *lemmatization*.

Regular Expressions for Detecting Word Patterns

- Many linguistic processing tasks involve pattern matching.
- Regular expressions (RE) give us a powerful and flexible method for describing the character patterns we are interested in.

Dr. Isma F. Siddiqui (isma.farah@faculty.muet.edu.pk)

Is stemming useful?

- For IR performance, some improvement (especially for smaller documents)
- May help a lot for some queries, but on average (across all queries) it doesn't help much (i.e. for some queries the results are worse)
  - Word sense disambiguation on query terms: *business* may be stemmed to *busy, saw (*the tool*) to see*
  - A truncated stem can be intelligible to users
  - Most studies for stemming for IR done for English (may help more for other languages)
  - The possibility of letting people interactively influence the stemming has not been studied much
- Since improvement is small, often IR engine usually don't use stemming
- More on this when we'll talk about IR

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Text Normalization

- Stemming
- Convert to lower case
- Identifying *non-standard words* including numbers, abbreviations, and dates, and mapping any such tokens to a special vocabulary.
  - For example, every decimal number could be mapped to a single token 0.0, and every acronym could be mapped to AAA. This keeps the vocabulary small and improves the accuracy of many language modeling tasks.
- Lemmatization
  - Make sure that the resulting form is a known word in a dictionary
  - WordNet lemmatizer only removes affixes if the resulting word is in its dictionary

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Lemmatization

- WordNet lemmatizer only removes affixes if the resulting word is in its dictionary
- The WordNet lemmatizer is a good choice if you want to compile the vocabulary of some texts and want a list of valid lemmas

Tokenization

- Divide text into units called tokens (words, numbers, punctuations) (page 124—136 Manning)
- What is a word?
  - Graphic word: string of continuous alpha numeric character surrounded by white space
    - *$22.50*
  - Main clue (in English) is the occurrence of whitespaces
  - Problems
    - Periods: usually remove punctuation but sometimes it's useful to keep periods (*Wash.* → *wash*)
    - Single apostrophes, contractions (*isn't, didn't, dog's*: for meaning extraction could be useful to have 2 separate forms: *is* + *n't* or *not*)
    - Hyphenation:
      - Sometime best a single word: *co-operate*
      - Sometime best as 2 separate words: *26-year-old, aluminum-export ban*

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Tokenization

- Whitespace often do not indicate a word break: sometime we may want to lump together words that are separated by a white space (whitespace?) but that we want to regard as a single word
  - San Francisco
  - The New York-New Heaven railroad
  - Wake up, work out
    - I couldn't *work* the answer *out*

Tokenization

- Tokenization turns out to be a far more difficult task than you might have expected. No single solution works well across-the-board, and we must decide what counts as a token depending on the application domain.

- When developing a tokenizer it helps to have access to raw text which has been manually tokenized, in order to compare the output of your tokenizer with high-quality (or "gold-standard") tokens.

Segmentation

- Word segmentation
  - For languages that do not put spaces between words
    - Chinese, Japanese, Korean, Thai, German (for compound nouns)
- Sentence segmentation
  - Divide text into sentences
  - Why?

- Sentence:
  - Something ending with a .. ?, ! (and sometime also :)
  - “You reminded me,” she remarked, “of your mother.”
    - Nested sentences
    - Note the .”

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Language Errors

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Speech Recognition and Text Correction

- Commonalities:
  - Both concerned with problem of accepting a string of symbols and mapping them to a sequence of progressively less likely words
  - Spelling correction: characters
  - Speech recognition: phones

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Spelling Error Patterns

- According to Damereau (1964) 80% of all misspelled words are caused by single-error misspellings which fall into the following categories (for the word **the**):
  - Insertion (**ther**)
  - Deletion (**th**)
  - Substitution (**thw**)
  - Transposition (**teh**)
- Because of this study, much subsequent research focused on the correction of single error misspellings

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Causes of Spelling Errors

- **Keyboard Based**
  - 83% novice and 51% overall were keyboard related errors
  - Immediately adjacent keys in the same row of the keyboard (50% of the novice substitutions, 31% of all substitutions)
- **Cognitive**
  - Phonetic seperate – separate
  - Homonym there – their

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Tagging

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	' or "
POS	Possessive ending	*'s*	"	Right quote	' or "
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	[, (, {, <
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	],), }, >
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	. ! ?
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	: ; ... – -
RP	Particle	*up, off*			

Figure 5.6 Penn Treebank part-of-speech tags (including punctuation).

Dr. Isma F. Siddiqui (isma.farah@faculty.muet.edu.pk)

Part Of Speech Tagging

- Annotate each word in a sentence with a part-of-speech marker.
- Lowest level of syntactic analysis.

John saw the saw and decided to take it to the table.
NNP VBD DT NN CC VBD TO VB PRP IN DT NN

- Useful for subsequent syntactic parsing and word sense disambiguation.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

English POS Tagsets

- Original Brown corpus used a large set of 87 POS tags.
- Most common in NLP today is the Penn Treebank set of 45 tags.
  - Tagset used in these slides.
  - Reduced from the Brown set for use in the context of a parsed corpus (i.e. treebank).
- The C5 tagset used for the British National Corpus (BNC) has 61 tags.

English Parts of Speech

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)

English Parts of Speech (cont.)

- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
  - Basic (DT) a, an, the
  - WH-determiner (WDT): which, that
- Coordinating Conjunction (CC): and, but, or,
- Particle (RP): off (took off), up (put up)

Ambiguity in POS Tagging: A Problem

- "Like" can be a verb or a preposition
  - I like/VBP candy.
  - Time flies like/IN an arrow.
- "Around" can be a preposition, particle, or adverb
  - I bought it at the shop around/IN the corner.
  - I never got around/RP to getting a car.
  - A new Prius costs around/RB $25K.

POS Tagging Process

- Usually assume a separate initial tokenization process that separates and/or disambiguates punctuation, including detecting sentence boundaries.
- Degree of ambiguity in English (based on Brown corpus)
  - 11.5% of word types are ambiguous.
  - 40% of word tokens are ambiguous.
- Average POS tagging disagreement amongst expert human judges for the Penn treebank was 3.5%
  - Based on correcting the output of an initial automated tagger, which was deemed to be more accurate than tagging from scratch.
- Baseline: Picking the most frequent tag for each specific word type gives about 90% accuracy
  - 93.7% if use model for unknown words for Penn Treebank tagset.

POS Tagging Approaches

- **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge.
- **Learning-Based**: Trained on human annotated corpora like the Penn Treebank.
  - **Statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)
  - **Rule learning**: Transformation Based Learning (TBL)
  - **Neural networks**: Recurrent networks like Long Short Term Memory (LSTMs)
- Generally, learning-based approaches have been found to be more effective overall, taking into account the total amount of human expertise and effort involved.

Dr. Isma F. Siddiqui *(isma.farah@faculty.muet.edu.pk)*

Useful Python Libraries for NLP

- **1. Natural Language Toolkit (NLTK)**
  - Link: https://www.nltk.org/
- **2. TextBlob**
  - Link: https://textblob.readthedocs.io/en/dev/
- **4. Gensim**
  - Link: https://github.com/RaRe-Technologies/gensim
- **6. polyglot**
  - Link: https://polyglot.readthedocs.io/en/latest/index.html
- **7. scikit–learn**
  - Link: https://scikit-learn.org/
- **8. Pattern**
  - Link: https://www.clips.uantwerpen.be/pages/pattern