Natural Language Processing - Week02

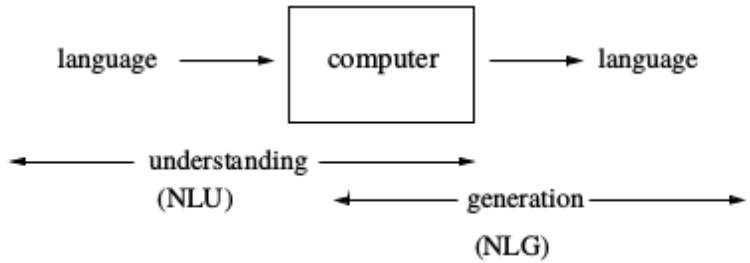
Assist. Prof. Dr. Tuğba YILDIZ

İSTANBUL BİLGİ UNIVERSITY
Department of Computer Engineering

March 1, 2017

Natural Language Processing (NLP)

- NLP is a field for computers to analyze, understand, and/or generating human language.
- computers using natural language as input and/or output
- concerns understanding and generating spoken and written text
- we will mostly concern with **written text** (not speech).



Natural Language Processing (NLP)

- Natural Language Understanding : Mapping the given input in the natural language into a useful representation.
- Natural Language Generation : Producing output in the natural language from some internal representation.

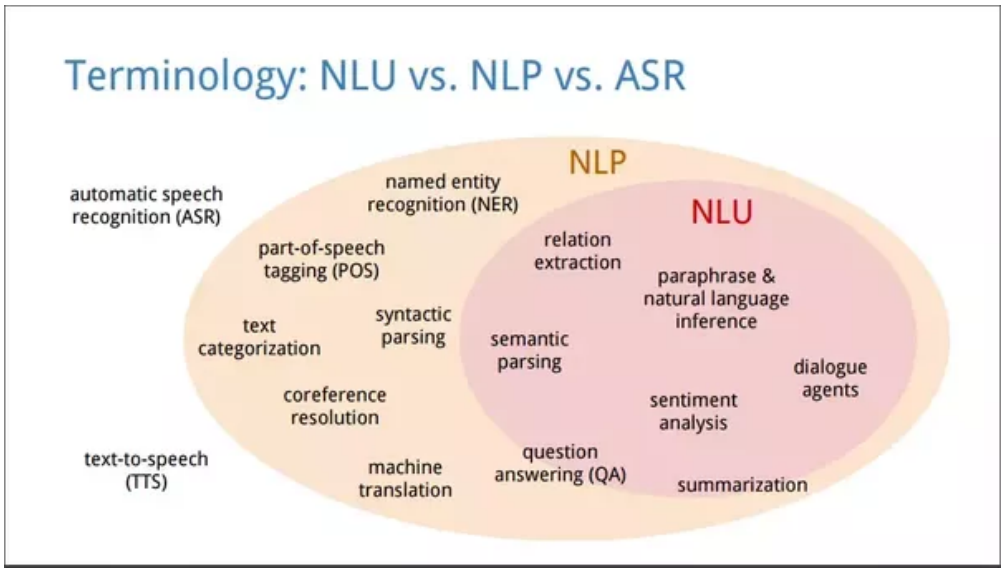


Fig.1 NLU in NLP [1]

[1] http://nlp.stanford.edu/ wcmac/papers/20140716-UNLU.pdf

Levels of Natural Language Processing (NLP)

- To process written text, we need:
  - Phonetics and Phonology - the study of linguistic sounds and their relations to words
    - Phonology : concerns units of speech
    - Phonetics : deals with how speech is organized
  - Morphology - knowledge of the meaningful components of words
  - Syntax - knowledge of the structural relationships between words
  - Semantics - knowledge of meaning
  - Pragmatics - knowledge of the relationship of meaning to the goals and intentions of the speaker (The meaning of the sentence depends on an understanding of the context and the speaker's intent.)
  - Discourse - knowledge about linguistic units larger than a single utterance

Levels of Natural Language Processing (NLP)

- Phonetics and Phonology - The study of linguistic sounds and their relations to words
- Turkish orthography is highly regular and a word's pronunciation is always completely identified by its spelling.

Turkish		IPA	English approximation	Turkish		IPA	English approximation
A	a	/a/	As *a* in ***f**a**t**h**e**r*	M	m	/m/	As *m* in ***m**a**n***
B	b	/b/	As *b* in ***b**o**y***	N	n	/n/	As *n* in ***n**i**c**e***
C	c	/d͡ʒ/	As *j* in ***j**o**y***	O	o	/o/	As *o* in *m**o**re*
Ç	ç	/t͡ʃ/	As *ch* in ***ch**a**i**r***	Ö	ö	/ø/	As *e* in ***le**a**r**n*, with lips rounded
D	d	/d/	As *d* in ***d**o**g***	P	p	/p/	As *p* in ***p**i**n***
E	e	/e/[1]	As *e* in ***r**e**d***	R	r	/ɾ/[2]	As *r* in ***r**i**n**g***
F	f	/f/	As *f* in ***f**a**r***	S	s	/s/	As *s* in ***s**o**n**g***
G	g	/g/, /ɟ/	As *g* in ***g**o**t***	Ş	ş	/ʃ/	As *sh* in ***sh**o**w***
Ğ	ğ	/ː/, /_/, /ˠ/	[3]	T	t	/t/	As *t* in ***t**i**c**k***
H	h	/h/	As *h* in ***h**o**t***	U	u	/u/	As *oo* in ***z**oo***
I	ı	/ɯ/	As *e* in *op**e**n*	Ü	ü	/y/	As *e* in ***n**e**w***
İ	i	/i/	As *ee* in ***f**ee**t***	V	v	/v/	As *v* in ***v**a**t***
J	j	/ʒ/	As *s* in *mea**s**ure*	Y	y	/j/	As *y* in ***y**e**s***
K	k	/k/, /c/	As *k* in ***k**i**t***	Z	z	/z/	As *z* in ***z**i**g**z**a**g***
L	l	/ɬ/, /l/	As *l* in ***l**o**v**e***				

Fig.2 It presents the Turkish letters, the sounds they correspond to in International Phonetic Alphabet and how these can be approximated more or less by an English speaker. [2]

[2] https://en.wikipedia.org/wiki/Turkish_alphabet

Levels of Natural Language Processing (NLP)

- Morpohology - The study of internal structures of words and how they can be modified
- Parsing complex words into their components

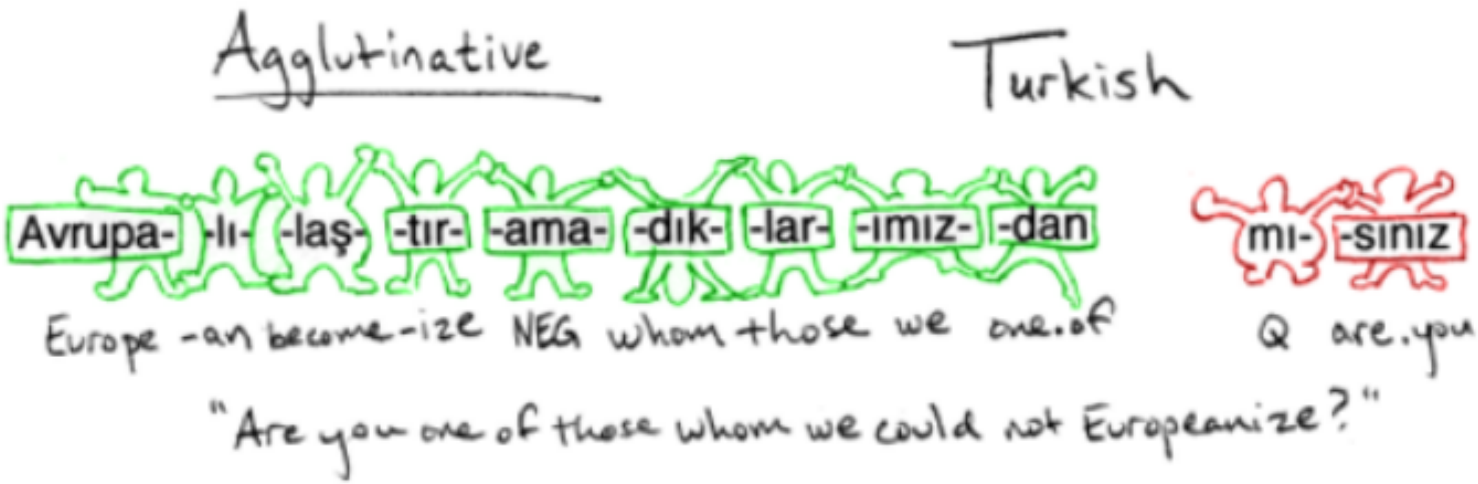


Fig.2 Moprhological Parsing [3]

[3] https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/NaturalLangu

Levels of Natural Language Processing (NLP)

- Morpohology - The study of internal structures of words and how they can be modified
- Relatively simple for English. But for some languages such as Turkish, it is more difficult.
  - uygarlaştıramadıklarımızdanmışsınızcasına
  - uygar-laş-tır-ama-dık-lar-ımız-dan-mış-sınız-casına
  - uygar +BEC +CAUS +NEGABLE +PPART +PL +P1PL +ABL +PAST +2PL +AsIf
  - "(behaving) as if you are among those whom we could not civilize/cause to become civilized"
    - +BEC is "become" in English
    - +CAUS is the causative voice marker on a verb
    - +PPART marks a past participle form
    - +P1PL is 1st person plural possessive marker
    - +2PL is 2nd person plural
    - +ABL is the ablative (from/among) case marker
    - +AsIf is a derivational marker that forms an adverb from a finite verb form
    - +NEGABLE is "not able" in English
- Inflectional and Derivational Morphology.
- Common tools: Finite-state transducers

Levels of Natural Language Processing (NLP)

- Morpohology - The study of internal structures of words and how they can be modified
- Part of Speech Tagging
  - Each word has a part-of-speech tag to describe its category.
  - Part-of-speech tag of a word is one of major word groups
  - open classes – noun, verb, adjective, adverb
  - closed classes – prepositions, determiners, conjuctions, pronouns, particples
  - POS Taggers try to find POS tags for the words.
  - duck is a verb or noun? (morphological analyzer cannot make decision).
  - A POS tagger may make that decision by looking the surrounding words.
  - Duck! (verb)
  - Duck is delicious for dinner. (noun)
- Inflectional and Derivational Morphology.
- Common tools: Finite-state transducers

Levels of Natural Language Processing (NLP)

- Syntax - The study of the structural relationships between words in a sentence
- Parsing - converting a flat input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.
- There are different parsing formalisms and algorithms.
- Most formalisms have two main components:
  - grammar – a declarative representation describing the syntactic structure of sentences in the language.
  - parser – an algorithm that analyzes the input and outputs its structural representation (its parse) consistent with the grammar specification.
- CFGs are in the center of many of the parsing mechanisms. But they are complemented by some additional features that make the formalism more suitable to handle natural languages.

Levels of Natural Language Processing (NLP)

- Syntax - The study of the structural relationships between words in a sentence
- Parsing complex words into their components

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound)))))))))))
    (VP (VBD were)
      (VP (VBN included)))
    (. .)))
```

Fig.2 Moprhological Parsing [4]

[4] https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/NaturalLangu

Levels of Natural Language Processing (NLP)

- Semantic - The study of the meaning of words, and how these combine to form the meanings of sentences
- Assigning meanings to the structures created by syntactic analysis.
- Mapping words and structures to particular domain objects in way consistent with our knowledge of the world.
- Semantic can play an import role in selecting among competing syntactic analyses and discarding illogical analyses.
- I robbed the bank – bank is a river bank or a financial institution
- We have to decide the formalisms which will be used in the meaning representation.

Levels of Natural Language Processing (NLP)

- Semantic - The study of the meaning of words, and how these combine to form the meanings of sentences
- Synonymy: fall - autumn
- Hypernymy - Hyponymy (is a): animal - dog
- Meronymy - Holonym (part-whole): finger - hand
- Antonymy: big - small

Levels of Natural Language Processing (NLP)

- Pragmatics - Social use of language
- 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
- Example:
  - Give me the salt!
  - Could you please give me the salt?

Levels of Natural Language Processing (NLP)

- Discourse - The study of linguistic units larger than a single statement
- John reads a book. He borrowed it from his friend.

Berlin (/bərˈlɪn/, German: [bɛɐ̯ˈliːn] (listen)) is the capital of Germany, and one of the 16 states of Germany. With a population of 3.5 million people,[4] Berlin is Germany's largest city. It is the second most populous city proper and the seventh most populous urban area in the European Union.[5] Located in northeastern Germany on the banks of River Spree, it is the center of the Berlin-Brandenburg Metropolitan Region, which has about 6 million residents from over 180 nations.[6][7][8][9] Due to its location in the European Plain, Berlin is influenced by a temperate seasonal climate. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes.[10]

Fig.2 NLP Levels with an Example [5]

[5] http://nlp.stanford.edu/ wcmac/papers/20140716-UNLU.pdf

Natural Language Processing (NLP)

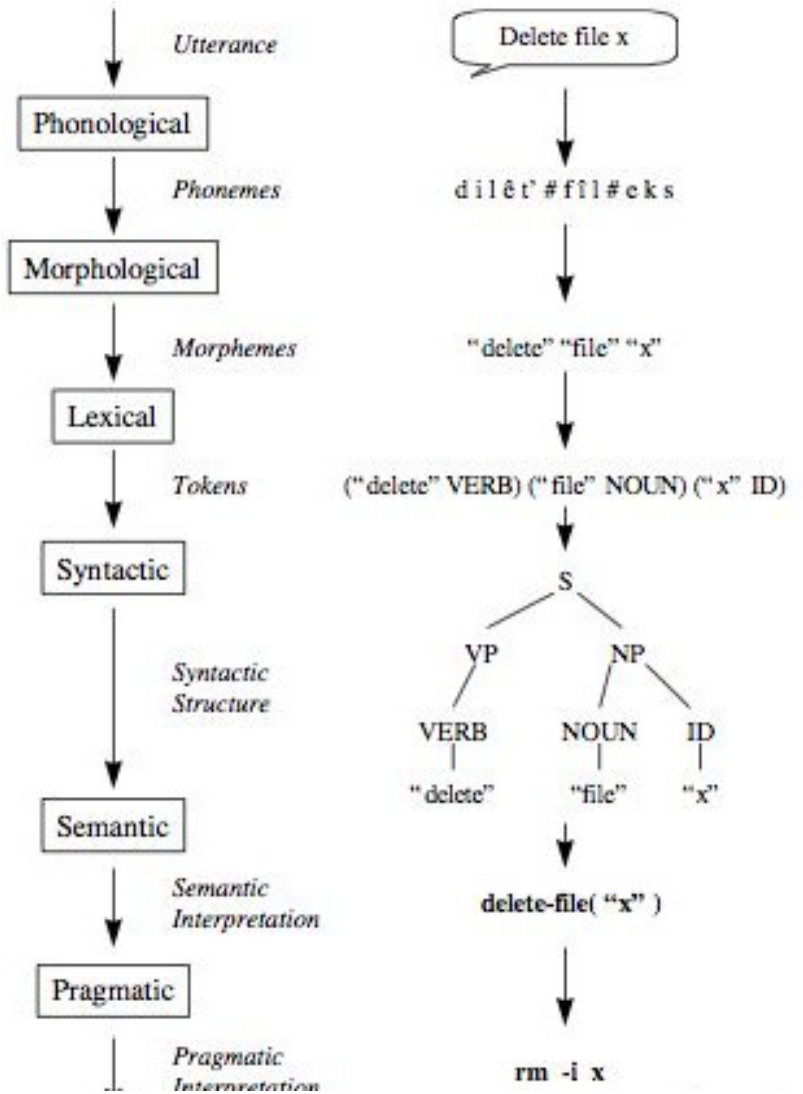


Fig.2 NLP Levels with an Example [6]

[6]

Natural Language Processing (NLP)

- Natural Language Understanding : Mapping the given input in the natural language into a useful representation.
- Different level of analysis required:
  - Words
  - Morphological Analysis : Morphologically analyzed words
  - Syntactic Analysis : Syntactic structure
  - Semantic analysis : Context-independent meaning representation
  - Discourse analysis : Final meaning representation

Natural Language Processing (NLP)

- Natural Language Generation : Producing output in the natural language from some internal representation.
- Different level of synthesis required:
  - Utterance
  - Utterance Planning : meaning representations for sentences
  - Sentence Planning and Lexical Choice : Syntactic structures of sentences with lexical choices
  - Sentence Generation : Morphologically analyzed words
  - Morphological Generation
  - Words

NLP Applications

- Applications:
  - automatic summarization
  - machine translation
  - named entity recognition
  - relationship extraction
  - sentiment analysis
  - speech recognition
  - topic segmentation
  - dialogue systems

NLP Applications: Dialogue Systems

- 2001: A Space Odyssey (directed by Stanley Kubrick)
  - -Dave: Open the pod bay doors, HAL.
  - -HAL: I'm sorry Dave, I'm afraid I can't do that.
  - -Dave: What's the problem?
  - -HAL: I think you know what the problem is just as well as I do.

Natural Language Processing (NLP)

- HAL must be able to
  - Phonetics and Phonology - recognize words from an audio signal and to generate audio signal from a sequence of words.
  - Morphology - produce and recognize words and other variations of individual words (I can't, doors)
    - ["I'm", "sorry", "Dave", "I'm", "afraid", "I", "can't", "."]
  - Syntax - use structural knowledge to propoerly string together the words that constitue its response.
    - "I'm I do, sorry that afraid Dave I'm can't."
  - Semantics - know meaning of the words
  - Syntax - use structural knowledge to propoerly string together the words that constitue its response.
    - "Yes, I'd like to hear it, HAL. Sing it for me!"

Natural Language Processing (NLP)

- HAL must be able to
  - Phonetics and Phonology - recognize words from an audio signal and to generate audio signal from a sequence of words.
  - Morphology - produce and recognize words and other variations of individual words (I can't, doors)
    - [“I'm”, “sorry”, “Dave”, “I'm”, “afraid”, “I”, “can't”, “.”]
  - Syntax - use structural knowledge to propoerly string together the words that constitue its response.
    - “I'm I do, sorry that afraid Dave I'm can't.”
  - Semantics - know meaning of the words
    - “Yes, I'd like to hear it, HAL. Sing it for me!”

Natural Language Processing (NLP)

- HAL must be able to
  - Phonetics and Phonology - recognize words from an audio signal and to generate audio signal from a sequence of words.
  - Morphology - produce and recognize words and other variations of individual words (I can't, doors)
    - [“I'm”, “sorry”, “Dave”, “I'm”, “afraid”, “I”, “can't”, “.”]
  - Syntax - use structural knowledge to propoerly string together the words that constitue its response.
    - “I'm I do, sorry that afraid Dave I'm can't.”
  - Semantics - know meaning of the words
    - “Yes, I'd like to hear it, HAL. Sing it for me!”

Natural Language Processing (NLP)

- HAL must be able to
  - Phonetics and Phonology - recognize words from an audio signal and to generate audio signal from a sequence of words.
  - Morphology - produce and recognize words and other variations of individual words (I can't, doors)
    - [“I'm”, “sorry”, “Dave”, “I'm”, “afraid”, “I”, “can't”, “.”]
  - Syntax - use structural knowledge to propoerly string together the words that constitue its response.
    - “I'm I do, sorry that afraid Dave I'm can't.”
  - Semantics - know meaning of the words
    - “Yes, I'd like to hear it, HAL. Sing it for me!”

Natural Language Processing (NLP)

- HAL must be able to
  - Phonetics and Phonology - recognize words from an audio signal and to generate audio signal from a sequence of words.
  - Morphology - produce and recognize words and other variations of individual words (I can't, doors)
    - [“I'm”, “sorry”, “Dave”, “I'm”, “afraid”, “I”, “can't”, “.”]
  - Syntax - use structural knowledge to propoerly string together the words that constitue its response.
    - “I'm I do, sorry that afraid Dave I'm can't.”
  - Semantics - know meaning of the words
    - “Yes, I'd like to hear it, HAL. Sing it for me!”

Natural Language Processing (NLP)

- HAL must be able to
  - Pragmatics - have knowledge about the kind of actions that speakers intend by their use of sentence.
    -Dave: Open the pod bay doors, HAL.
    -HAL: No or No, I wonnt' open the door. (-)
    -HAL: I'm sorry Dave, I'm afraid I can't do that. (+)
  - Discourse - need to examine the earlier questions that were asked.
    -Dave: Yes, I'd like to hear it, HAL. Sing it for me.
    -HAL: It's called "Daisy." Daisy, Daisy, give me your answer, do. I'm half crazy, all for the love of you. It won't be a stylish marriage.

Natural Language Processing (NLP)

- HAL must be able to
  - Pragmatics - have knowledge about the kind of actions that speakers intend by their use of sentence.
    -Dave: Open the pod bay doors, HAL.
    -HAL: No or No, I wonnt' open the door. (-)
    -HAL: I'm sorry Dave, I'm afraid I can't do that. (+)
  - Discourse - need to examine the earlier questions that were asked.
    -Dave: Yes, I'd like to hear it, HAL. Sing it for me.
    -HAL: It's called "Daisy." Daisy, Daisy, give me your answer, do. I'm half crazy, all for the love of you. It won't be a stylish marriage.

Why NLP is Hard?

- Ambiguity is at all levels of analysis:
  - Phonetics and phonology : "I scream" vs. "ice cream" / "recognize speech" vs. "wreck a nice beach"
  - Morphology : unionized = union + ized? / un + ionized?
  - Syntax:
    - John saw the man on the mountain with a telescope.
    - Who is on the mountain? John, the man, or both? Who has the telescope? John, the man, or the mountain?
  - Semantics: The astronomer loves the star. (star in the sky / celebrity)
  - Pragmatic:
    - "I just came from New York."
    - Would you like to go to New York today?
    - Would you like to go to Boston today?
    - Boy, you look tired.
  - Discourse: John and Mary took two trips around France. They were both wonderful.

Ambiguity

- I made her duck.
  - How many different interpretations does this sentence have?
  - What are the reasons for the ambiguity?
  - The categories of knowledge of language can be thought of as ambiguity resolving components.
  - How can each ambiguous piece be resolved?

Ambiguity (cont)

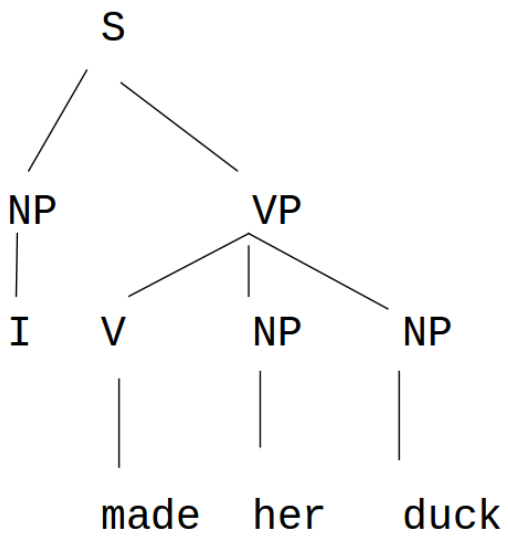
- Some interpretations of : I made her duck.
  1. I cooked duck for her.
  2. I cooked duck belonging to her.
  3. I created a toy duck which she owns.
  4. I caused her to quickly lower her head or body.
  5. I used magic and turned her into a duck.
- duck – morphologically and syntactically ambiguous: noun or verb.
- her – syntactically ambiguous: dative or possessive.
- make – semantically ambiguous: cook or create.
- make – syntactically ambiguous

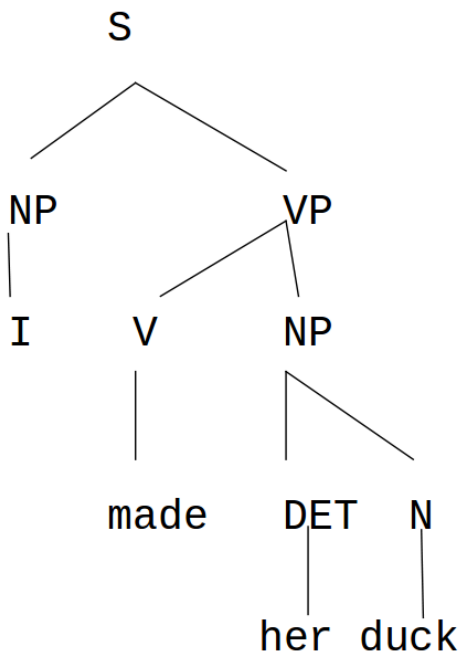
Ambiguity (cont)

- I made her duck.
- part-of-speech tagging - Deciding whether duck is verb or noun.
- word-sense disambiguation - Deciding whether make is create or cook.
- lexical disambiguation - Resolution of part-of-speech and word-sense ambiguities are two important kinds of lexical disambiguation.
- syntactic ambiguity - her duck is an example of syntactic ambiguity, and can be addressed by probabilistic parsing.

Ambiguity (cont)

- I made her duck.





Example: Ambiguity in Turkish

- masalı
  1. masal +Noun+A3sg+Pnon+Acc (= the story)
  2. masal +Noun+A3sg+P3sg+Nom (= his story)
  3. masa +Noun+A3sg+Pnon+NomaDB+Adj+With (= with tables)

Models and Algorithms

- state machines, rule systems, logic, probabilistic models, and vector-space models.
  - state machines (deterministic and non-deterministic finite-state automata and finite-state transducers)
  - regular grammars and regular relations, context-free grammars
  - first order logic, also known as the predicate calculus
  - hidden Markov models or HMMs
  - vector-space models, based on linear algebra
  - machine learning tools like classifiers (decision trees, support vector machines, logistic regression, etc.)

NLP - An Interdisciplinary Field

- NLP borrows techniques and insights from several disciplines.
  - Linguistics: How do words form phrases and sentences? What constraints the possible meaning for a sentence?
  - Computational Linguistics: How is the structure of sentences are identified? How can knowledge and reasoning be modeled?
  - Computer Science: Algorithms for automatons, parsers.
  - Engineering: Stochastic techniques for ambiguity resolution.
  - Psychology: What linguistic constructions are easy or difficult for people to learn to use?
  - Philosophy: What is the meaning, and how do words and sentences acquire it?

Regular Expressions and Automata

- Regular Expressions (RE): the standard notation for characterizing text sequences, a language for specifying text search strings
- RE is used for specifying text strings in situations like
  - Web-search
  - information retrieval
  - word processing
  - computing frequency from text

Regular Expressions and Automata

- RE is a formula in a special language that is used for specifying simple classes of strings.
- A string is a sequence of symbols, any sequence of alphanumeric character
- RE search requires a pattern that we want to search in text.

Regular Expressions and Automata

- How can we search for any of these?
  - woodchucks

RE	Example Patterns Matched
/woodchucks/	"interesting links to woodchucks and lemurs"
/a/	"Mary Ann stopped by Mona's"
/Claire_says,/	"Dagmar, my gift please," Claire says,"
/DOROTHY/	"SURRENDER DOROTHY"
/!/	"You've left the burglar behind again!" said No

Regular Expressions

- woodchuck
- Woodchuck
- Letters inside square brackets

RE	Match	Example Patterns
/[wW]oodchuck/	Woodchuck or woodchuck	"Woodchuck"
/[abc]/	'a', 'b', *or* 'c'	"In uomini, in soldati"
/[1234567890]/	any digit	"plenty of 7 to 5"

Regular Expressions

- Ranges

RE	Match	Example Patterns Matched
/[A-Z]/	an uppercase letter	"we should call it 'Drenched Blossoms'"
/[a-z]/	a lowercase letter	"my beans were impatient to be hoed!"
/[0-9]/	a single digit	"Chapter 1: Down the Rabbit Hole"

Regular Expressions

- Negations
- Carat means negation only when first in [].

RE	Match (single characters)	Example Patterns Matched
[^A-Z]	not an uppercase letter	"Oyfn pripetchik"
[^Ss]	neither 'S' nor 's'	"I have no exquisite reason for't"
[^\.]	not a period	"our resident Djinn"
[e^]	either 'e' or '^'	"look up ^ now"
a^b	the pattern 'a^b'	"look up a^ b now"

Regular Expressions

- Optional previous character

RE	Match	Example Patterns Matched
`woodchucks?`	woodchuck or woodchucks	"woodchuck"
`colou?r`	color or colour	"colour"

Regular Expressions

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
groundhog|woodchuck	
yours|mine	yours mine
a|b|c	= [abc]
[gG]roundhog|[Ww]oodchuck	

Regular Expressions

- Stephen C Kleene (Kleene *, Kleene +)

colou?r	optional previous char	color colour
ba!*	0 or more of previous char	ba ba! ba!! ba!!! ba!!!!
ba!+	1 or more of previous char	ba! ba!! ba!!! ba!!!!
beg.n	1 any char	begin begun beg1n beg.n

Regular Expressions

- {n} : The preceding item is matched exactly n times.
- {n,} : The preceding item is matched n or more times
- {,m} :The preceding item is matched at most m times.
- {n,m}: The preceding item is matched at least n times, but not more than m times.
- for the special characters use backslash character

Regular Expressions

- Anchors

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
\.$	The end.
.$	The end? The end!

Regular Expressions

- Example:
  - Find the words which has 4 characters
  - Find the words which has 4 characters

Regular Expressions

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)
- Reducing the error rate :
- Increasing accuracy or precision (minimizing false positives)
- Increasing coverage or recall (minimizing false negatives).

Regular Expressions

- Regular expressions play a surprisingly large role
- Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
- But regular expressions are used as features in the classifiers
- Can be very useful in capturing generalizations

Finite State Automaton

- Regular Expressions (RE): the standard notation for characterizing text sequences
- RE is used for specifying text strings in situations like
  - Web-search
  - information retrieval
  - word processing
  - computing frequency from text

References I