NLP Overview

정윤경
성균관대학교
소프트웨어학과

본 자료는 2013년도 GAMNLP 워크샵 튜토리얼 (Noriko Tomuro 교수 작성)을 토대로 작성됨

강사 및 조교

- 정윤경 (aimecca@skku.edu)
- 학력
  - 성균관대 정보공학과 학/석사
  - 미국 North Carolina State University CS 박사 (인공지능 전공)
- 경력
  - LG전자 연구원
  - 삼성전자 종합기술원 전문연구원 (AI 과제 기획, interactive storytelling)
  - IT University of Copenhagen, Denmark, post-doc (Game AI, data analysis)
  - 성균관대 소프트웨어대학 조교수

- 조교: 김유진, 이석범

운영 방식

- 내용
  - NLP Overview
  - Syntax Processing
  - Semantic Processing
  - Sentence Processing
  - Natural Language Understanding

- 수업 방식
  - 이론 5시간, 실습 3시간
  - 50분 수업, 15분 휴식

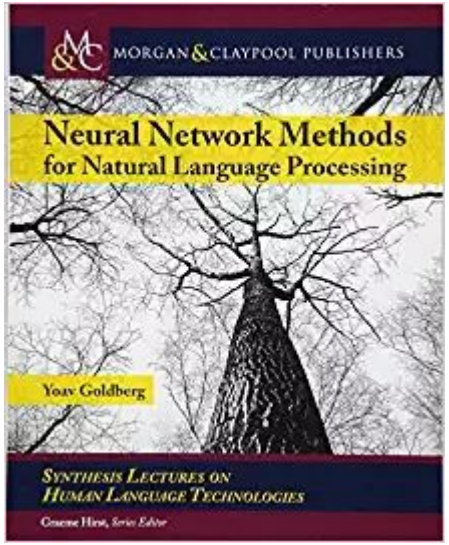
무료 참고 자료

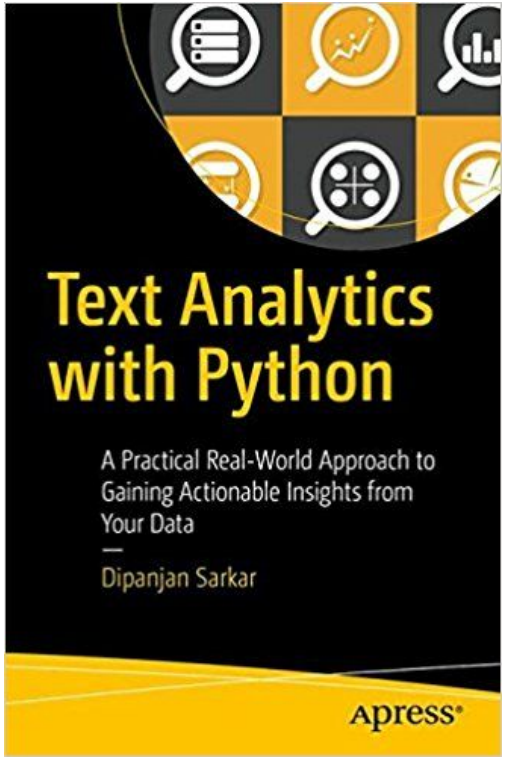
- 한국어
  - 점프 투 파이썬: https://wikidocs.net/book/1
  - 코딩도장: https://dojang.io/course/view.php?id=3
  - NLP 정리: http://docs.likejazz.com/deep-learning-for-nlp/

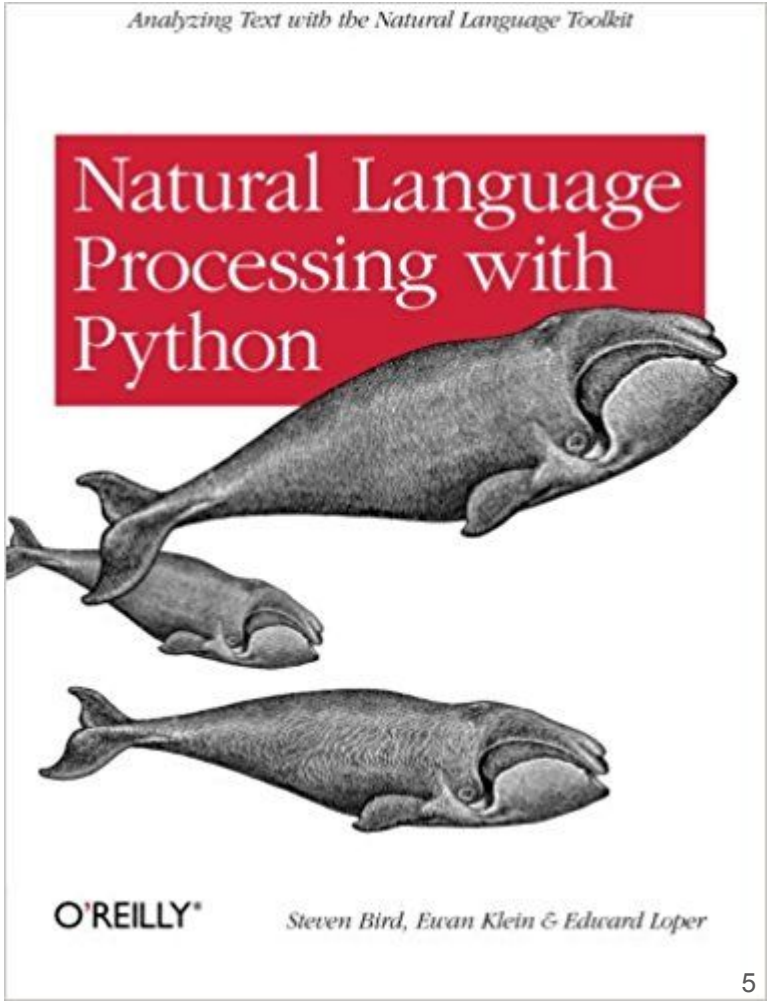
- 영어
  - Automate the boring stuff: https://automatetheboringstuff.com/
  - How to think like a computer scientist: http://interactivepython.org/runestone/static/thinkcspy/index.htm
  - Spacy: https://spacy.io/usage/spacy-101

NLP 참고 문헌

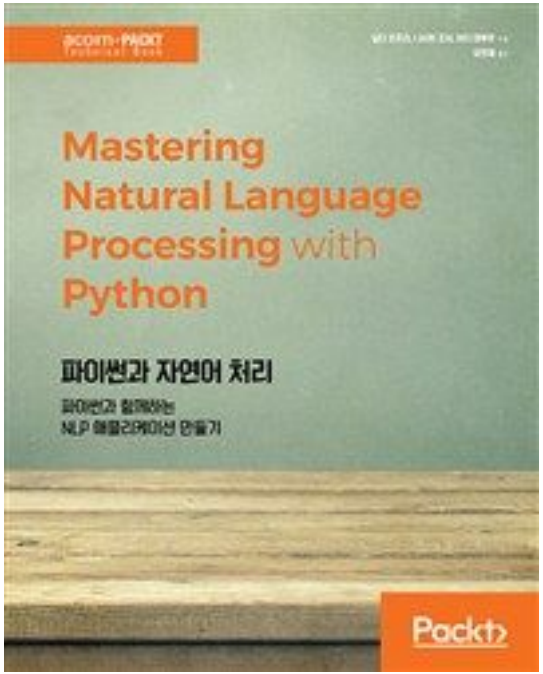
http://www.nltk.org/

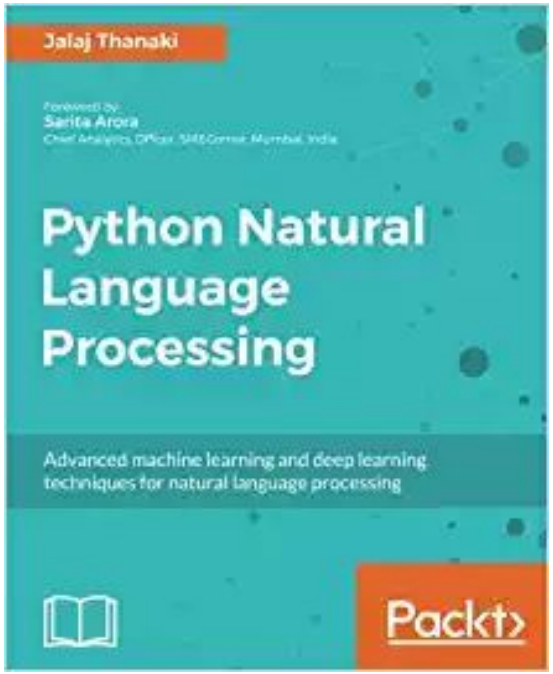


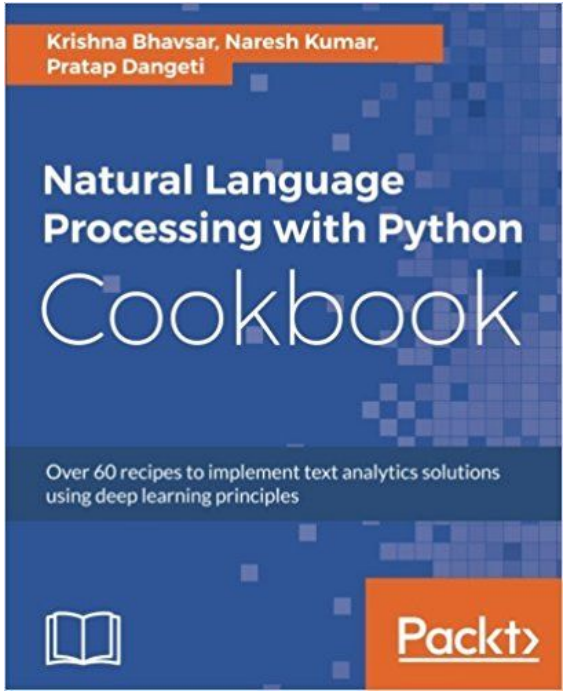




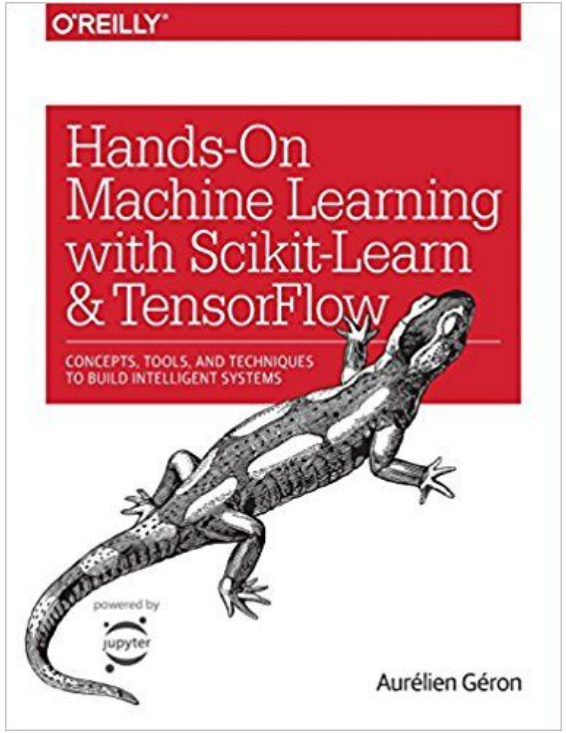
NLP 기타

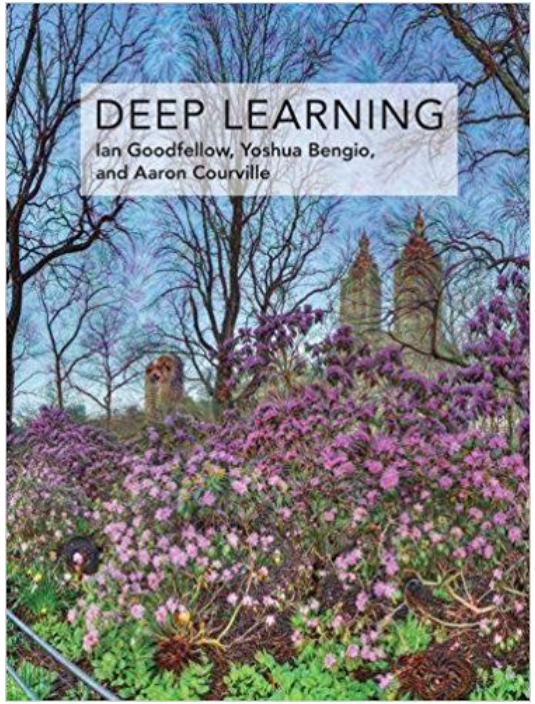

Mastering
Natural Language
Processing with
Python
파이썬과 자연어 처리
Packt

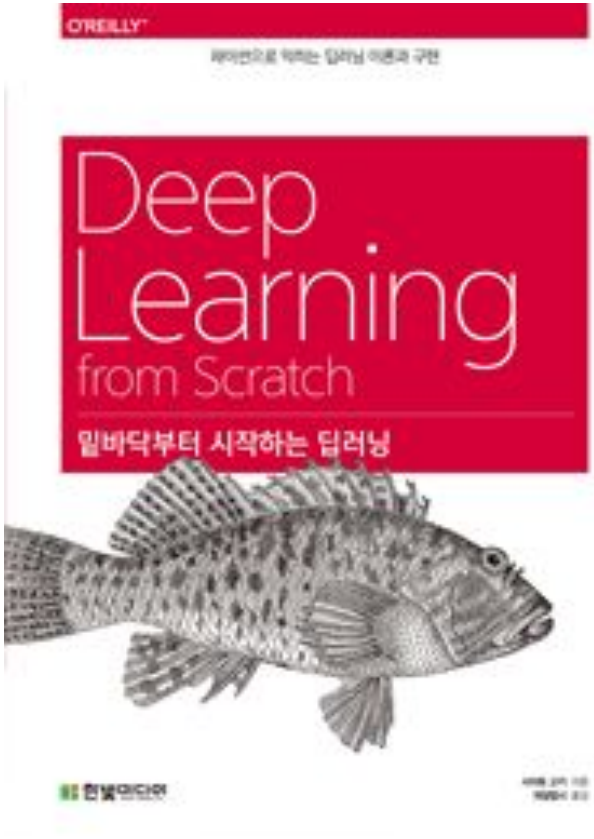

Jalaj Thanaki
Sarita Arora
Python Natural
Language
Processing
Advanced machine learning and deep learning
techniques for natural language processing
Packt


Krishna Bhavsar, Naresh Kumar,
Pratap Dangeti
Natural Language
Processing with Python
Cookbook
Over 60 recipes to implement text analytics solutions
using deep learning principles
Packt

ML and Deep Learning





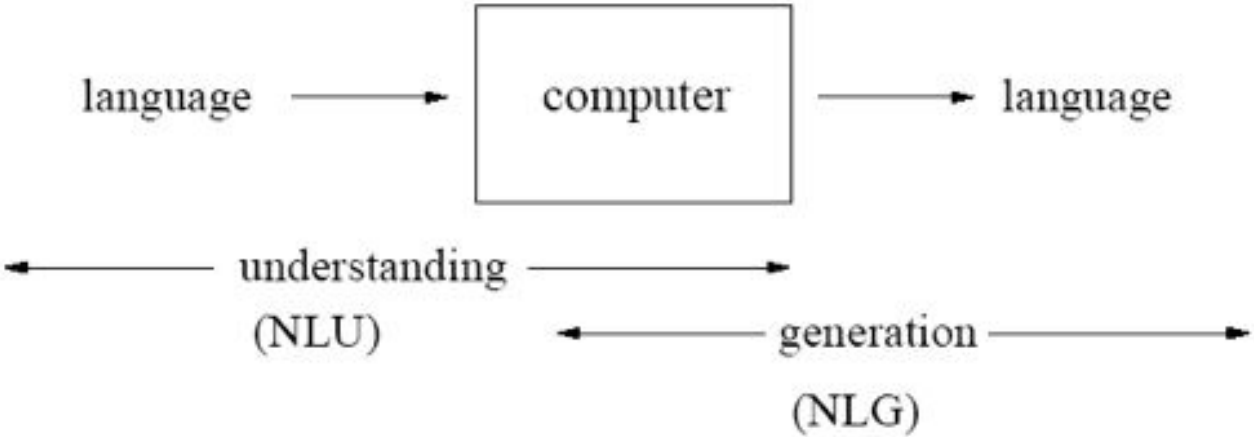


NLP 연구 접근 방법

- 규칙 기반
- 통계적 기반
- 딥러닝 기반

What is NLP?

Natural Language Processing (NLP) is a field in Artificial Intelligence (AI) devoted to creating computers that use natural language as input and/or output.



NLP is an Interdisciplinary Field

- Linguistics: NLP is also called ”Computational Linguistics”
- Psychology
- Mathematics and Statistics
- Information Theory
- Computer Science, AI

Machines that can speak

HAL 9000 in “2001: A Space Odyssey”

C3PO in Star Wars

KITT in Knight Rider

NLP is AI-complete

- The most difficult problems in AI
- Language is ambiguous
- Requires world knowledge and logical reasoning

Formal Linguistic Analysis

- Phonology: speech audio signal to phonemes
- Morphology
  - Inflection (e.g. “I”, “my”, “me”; “eat”, “eats”, “ate”, “eaten”)
  - Derivation (e.g. “teach”, “teacher”, “nominate”, “nominee”)
- Syntax:
  - Part-of-speech (noun, verb, adjective, preposition, etc.)
  - Phrase structure (e.g. noun phrase, verb phrase)
- Semantics: meaning of a word (e.g. “book” as a bound volume or an accounting ledger) or a sentence
- Discourse: meaning and inter-relation between sentences (and dialogues)
- Pragmatic Analysis: purposeful use of sentences in situations

Source: Joyce Choi, CSE 842, Michigan State University

Ways to Form Words

- Inflection: new forms of the same word (usually in the same class)
  - Tense, number, mood, voice marking in verbs
  - Number, gender marking in nominals
  - Comparison of adjectives
- Derivation: yield different words in different class
  - Deverbal nominals: 동사 파생 명사
  - Denominal adjectives and verbs: 명사 파생 동사
- Compounding: new words out of two or more other words
  - Noun-noun compounding (e.g., doghouse)
- Cliticization: combine a word with a clitic (which acts syntactically like a word but in a reduced form, e.g., I've)

Morphology

- The study of how words are composed of morphemes (the smallest meaning-bearing units of a language)
- Two broad classes of morphemes:
  - Stems: “main” morpheme of the word, supplying meaning
  - Affixes: Bits and pieces that combine with stems to modify their meanings and grammatical functions (prefixes, suffixes, circumfixes, infixes)
    - Unlike
    - Trying
  - Multiple affixes

    Unreadable

Morphological Analysis Tools

- Porter stemmer
  - A simple approach: just hack off the end of the word!
  - Does NOT convert a word to its base form
  - Frequently used in Information Retrieval, but results are pretty ugly!

Original **********
Rudolph Agnew , 55 years old and former chairman of
Consolidated Gold Fields PLC , was named a **nonexecutive** director of
this British **industrial conglomerate** . A form of **asbestos once** used to
make Kent **cigarette filters has** caused a high **percentage** of cancer
deaths among a group of workers **exposed** to it more than 30 years ago ,

Results **********
Rudolph Agnew , 55 year old and former chairman of
Consolid Gold Field PLC , wa name a **nonexecut** director of
thi British **industri conglomer** . A form of **asbesto onc** use to
make Kent **cigarett filter ha caus** a high **percentag** of cancer
death among a group of worker **expos** to it more than 30 year ago ,

Source: Marti Hearst, i256, at UC Berkeley

Morphological Analysis Tools

- WordNet's morphy()
  Use transformation rules and built-in lookup tables
- Very sophisticated programs have been developed
  Best known: PCKimmo
  Commercial versions: inXight's LinguistX

POS	Suffix	Ending
NOUN	"s"	""
NOUN	"ses"	"s"
NOUN	"xes"	"x"
NOUN	"zes"	"z"
NOUN	"ches"	"ch"
NOUN	"shes"	"sh"
NOUN	"men"	"man"
NOUN	"ies"	"y"
VERB	"s"	""
VERB	"ies"	"y"
VERB	"es"	"e"
VERB	"es"	""
VERB	"ed"	"e"
VERB	"ed"	""
VERB	"ing"	"e"
VERB	"ing"	""
ADJ	"er"	""
ADJ	"est"	""
ADJ	"er"	"e"
ADJ	"est"	"e"

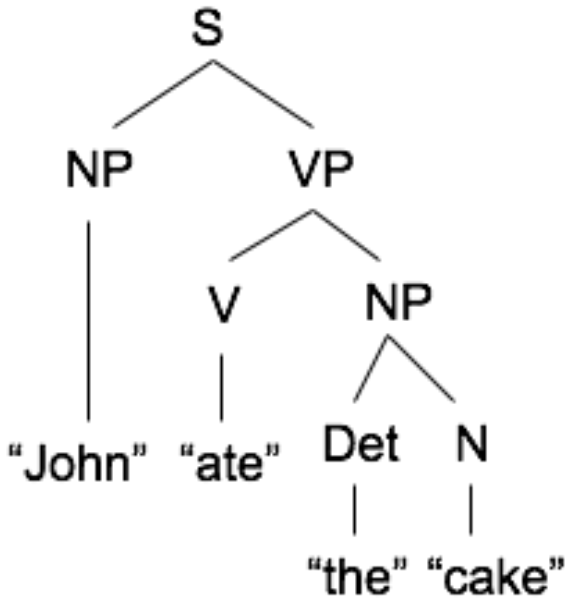
Source: Marti Hearst, i256, at UC Berkeley

Syntactic Parsing

“John ate the cake”

Grammar

R0: S → NP VP
R1: NP → Det N
R2: VP → VG NP
R3: VG → V
R4: NP → "John"
R5: V → "ate"
R6: Det → "the"
R7: N → "cake"



Semantic Analysis

- Derive the meaning of a sentence.
- Often applied on the result of syntactic analysis.

  "John ate the cake."

  NP V NP

  ((action INGEST) ; syntactic verb

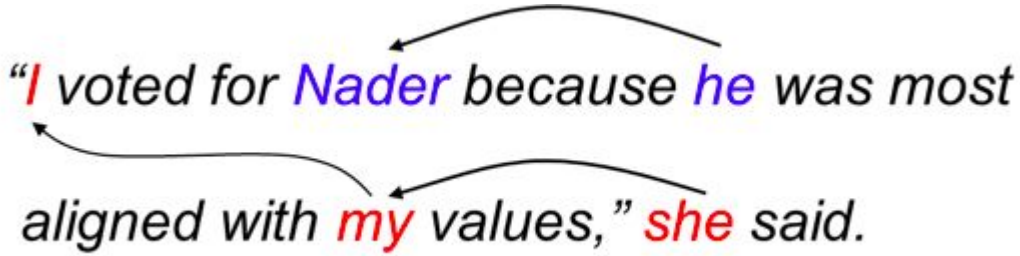
  (actor JOHN-01) ; syntactic subj

  (object FOOD)) ; syntactic obj
- To do semantic analysis, we need a (semantic) dictionary (e.g. WordNet, http://www.cogsci.princeton.edu/~wn/).

Discourse Analysis

- Go beyond just one sentence – several sentences paragraph, ‘block of text’
- Analyze relations between sentences, including:

1. Anaphora (e.g. pronouns such as “he”, “she”, “it”, “they”) & Co-reference resolution.



Discourse Analysis

2. Quantification (e.g. “All participants had a cake for desert.” – How many cakes were there?)

3. Entailment in the real world model (by inference)
e.g. “He was snoring.” =(entails)=> He was sleeping.

4. Topic shift – segmentation of blocks; physical markings (e.g. paragraph break, keywords) are not present all the time.

Pragmatics Analysis

- Views language as social interaction between agents (i.e., social science) rather than descriptive texts.
- Analyses include:
  - World knowledge
  - Speech act: an utterance that has performative function in language and communication
    Speaker X: "We should leave for the show or else we'll be late." (request, suggestion)
    Speaker Y: "I am not ready yet." (statement, rejection)

Speech act (John Searle)

- **Representatives** commit a speaker to the truth of an expressed proposition.
  Paradigm cases: asserting, stating, concluding, boasting, describing, suggesting.
  I am a great singer.
  Bill was an accountant.
- **Commissives** commit a speaker to some future action.
  Paradigm cases: promising, pledging, threatening, vowing, offering.
  I am going to leave you.
  I'll call you tonight.
- **Directives** are used by a speaker who attempts to get the addressee to carry out an action.
  Paradigm cases: requesting, advsing, commanding, challenging, inviting, daring, entreating.
  You'd better tidy up that mess.
  Sit down.

John Searle, *Speech Acts*, Cambridge University Press 1969

Speech act

- **Declarations** affect an immediate change of affairs.
  Paradigm cases: declaring, baptising, resigning, firing from employment, hiring, arresting.
  We find the defendant guilty.
  I resign.
- **Expressives** express some sort of psychological state.
  Paradigm cases: greeting, thanking, apologising, complaining, congratulating.
  This beer is disgusting.
  I'm sorry to hear that.

John Searle, *Speech Acts*, Cambridge University Press 1969

Demo

http://text-processing.com/demo/

http://text-processing.com/demo/sentiment/

http://textanalysisonline.com/nltk-pos-tagging

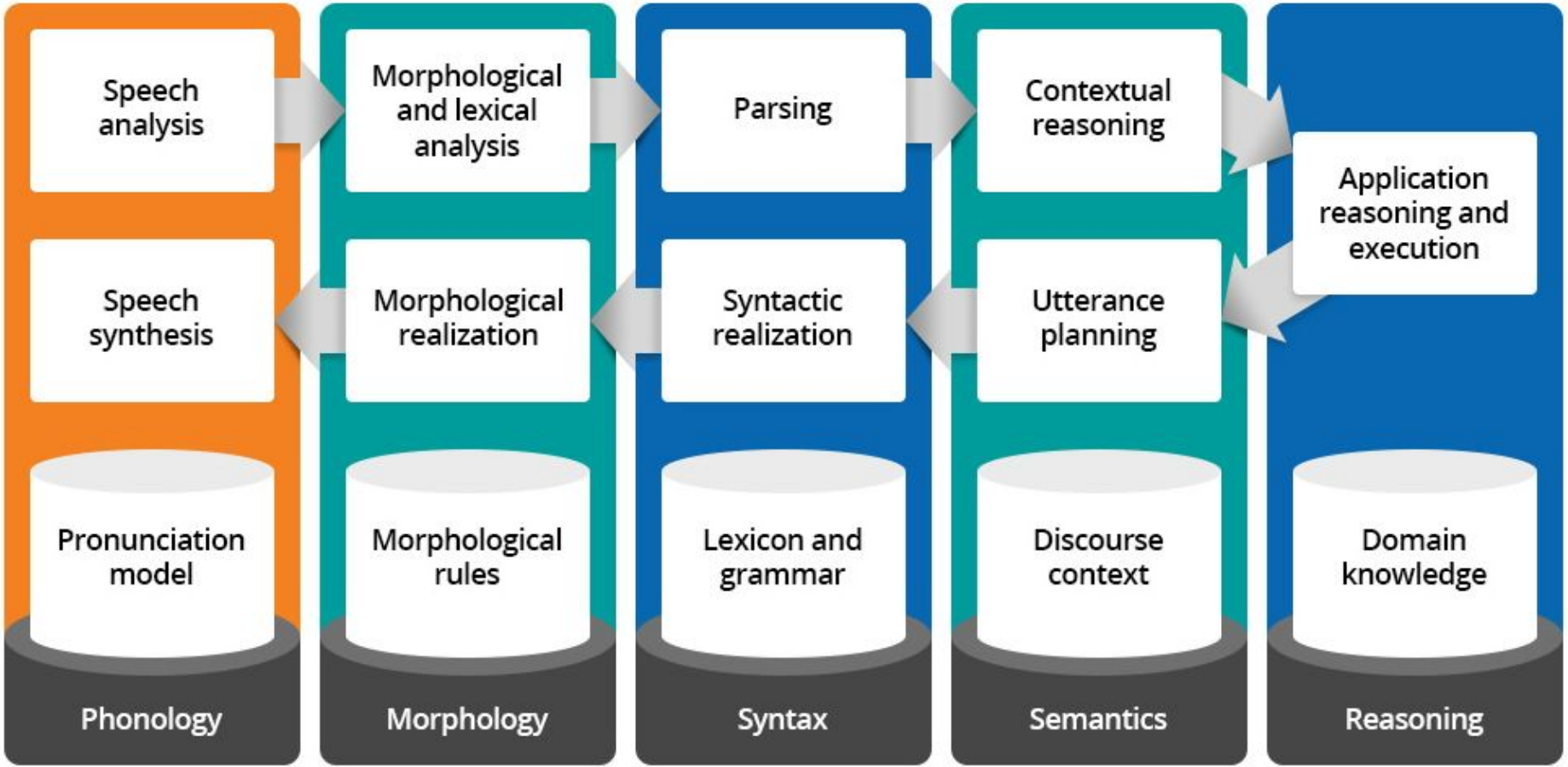
Stanford Parser: http://nlp.stanford.edu:8080/parser/

http://www.conversational-technologies.com/nldemos/nlDemos.html

Why is NLP so hard..?

because of inherent *ambiguity*

"Get the cat with the gloves."



http://mediterra-soft.com/research.html

Ambiguity - Phonetics

I mate or duck
I'm eight or duck
Eye maid; her duck
Aye mate, her duck
I maid her duck
I'm aid her duck
I mate her duck
I'm ate her duck
I'm ate or duck
I mate or duck

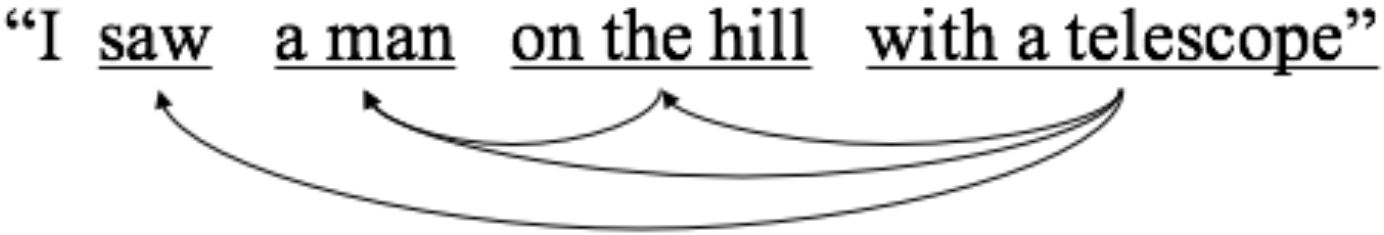
Sound like
"I made her duck"

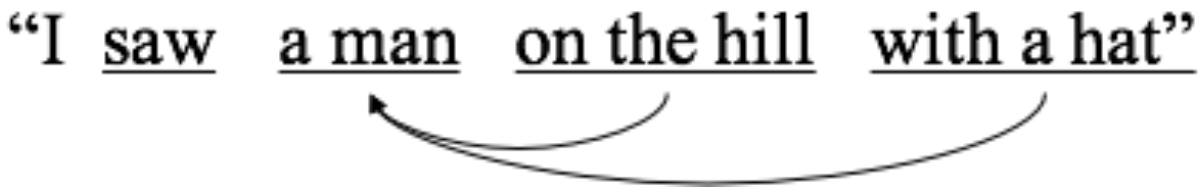
Source: Jurafsky & Martin "Speech and Language Processing"

Ambiguity - Lexical

- Lexical category (part-of-speech)
  - “duck” as a noun or a verb
- Lexical Semantics (word meaning)
  - “duck” as an animal or a plaster duck statue
- Compound nouns
  e.g. “dog food”, “Intelligent design scores …”

Ambiguity - Syntax





Ambiguity - Semantics

The car hit the pole while it was moving.

This can mean

The car, while moving, hit the pole. (a)

The car hit the pole while the pole was moving. (b)

(a) is preferred to (b) because we have a model of the world that helps us to distinguish what is logical (or possible) from what is not.

Ambiguity - Semantics

Kent Bach (1982) explores the intriguing case of:

I love you too.

This can mean (at least) one of four distinct things:

I love you (just like you love me)
I love you (just like someone else does)
I love you (and I love someone else)
I love you (as well as bearing some other relationship (i.e. liking) to you)

Formal Linguistic Analysis

- Phonology: speech audio signal to phonemes
- Morphology
  - Inflection (e.g. “I”, “my”, “me”; “eat”, “eats”, “ate”, “eaten”)
  - Derivation (e.g. “teach”, “teacher”, “nominate”, “nominee”)
- Syntax:
  - Part-of-speech (noun, verb, adjective, preposition, etc.)
  - Phrase structure (e.g. noun phrase, verb phrase)
- Semantics: meaning of a word (e.g. “book” as a bound volume or an accounting ledger) or a sentence
- Discourse: meaning and inter-relation between sentences (and dialogues)

Source: Joyce Choi, CSE 842, Michigan State University

Compounding Factors

- Ambiguity at different levels compound the difficulties of NLP.
- Engineering of NLP systems is also hard … because:
  - Huge amount of data resources needed (e.g. grammar, dictionary, documents to extract statistics from)
  - Computational complexity (intractable) of analyzing a sentence

NLP 응용

- Conversational AI, Intelligent Assistant
- Automatic summarization
- Question Answering
- Sentiment analysis
- News Article Generation
- Information Retrieval
- Machine Translation
- Text Summarization
- Document Analysis and Classification

Early Conversational Programs

- ELIZA (by Joseph Weizenbaum), 1966
  - A psychotherapist
  - No real understanding; simple pattern-matching to respond to user input -> canned responses

```
>> Hello.
How do you do.  Please state your problem.
>> Men are all alike.
In what way ?
>> They're alwas bugging us about something or other.
I'm not sure I understand you fully.
>> They're always bugging us about something or other.
Can you think of a specific example ?
>> Well, my boyfriend made me come here.
Your boyfriend made you come here  ?
>> He says I'm depressed much of the time.
I am sorry to hear that you are depressed.
```

Turing Test (Alan Turing, 1950)

test of a machine's capability to perform human-like conversation

A human judge engages in a natural language conversation with two other parties, one a human and the other a machine; if the judge cannot reliably tell which is which, then the machine is said to pass the test.

심심이

집단지성이 직접 작성한 대화 문장

물어보는 말 - 대답하는 말 쌍을 입력

중국어 방 (John Searle, 1980)

- Imagine that a man who does not speak Chinese sits in a room and is passed Chinese symbols through a slot in the door. To him, the symbols are just so many squiggles and squoggles. But he reads an English-language rule book that tells him how to manipulate the symbols and which ones to send back out.

튜링 테스트로는 기계의 인공지능 여부를 판정할 수 없다는 것을 논증하기 위해 고안한 사고실험.

https://debategraph.org/Details.aspx?nid=219

IBM Watson

Q&A system

Games with Conversational Agents

- Façade

  Mostly pattern matching, with some discourse acts.

```
(defrule positional_Is
  (template (tor am are is seem seems
sound sounds look looks))
   => (assert (iIs ?startpos
?endpos)))
```

  The above rule says if you see any words that approximate the verb to be, assert into working memory a new fact of iIs at the position found.

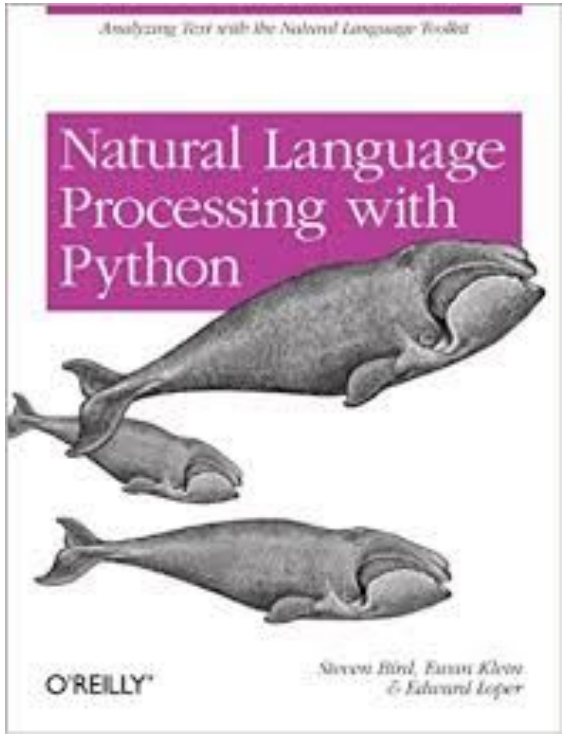
Games with Conversational Agents

- The Restaurant game
  a restaurant simulation where
  you play as a customer or waitress

교재

Natural language processing with python

https://pdfs.semanticscholar.org/3673/bccde93025e05431a2bcac4e8ff18c9c273a.pdf



syllabus

- **NLP Overview**
- Syntax Processing
- Semantic Processing
- Sentence Processing
- Natural Language Understanding

- Language Syntax and Structure
- Python review
- Corpora and Lexical Resources
- Reading from Web, pdf, word, csv
- Unicode
- Regular Expression
- Small letter conversion, punctuation

syllabus

- NLP Overview
- **Syntax Processing**
- Semantic Processing
- Sentence Processing
- Natural Language Understanding

- Sentence and Word Tokenization
- Removal of stopwords
- Stemming and Lemmatization
- POS Tagging
- n-gram
- Feature Extraction
- tf-idf, one-hot coding
- BOW
- Sentiment Analysis
- Text Classification: spam detection, genre categorization

syllabus

- NLP Overview
- Syntax Processing
- **Semantic Processing**
- Sentence Processing
- Natural Language Understanding

- Chunking
- NER
- Word Embedding
- Word2Vec, skip gram
- ambiguity
- Relation Extraction
- Pronoun Resolution

syllabus

- NLP Overview
- Syntax Processing
- Semantic Processing
- **Sentence Processing**
- Natural Language Understanding

- Syntax
- 문장 구조
- Context-Free Grammar
- Parsing with CFG
- Dependencies

syllabus

- NLP Overview
- Syntax Processing
- Semantic Processing
- Sentence Processing
- **Natural Language Understanding**

- Semantic Representation:
Propositional Logic, First-order Logic
- Discourse
- NLP응용
- document analysis, NLG
- 한글 관련 라이브러리

강의 수강을 위한 기본 지식

- 파이썬에 대한 이해: list, 문자열, dictionary, list comprehension
  - E.g., long_words = [w for w in V if len(w) > 15]
- Plot graph
- 단순한 통계 지식
- 영문법

Data Mining Research Process

01 Data Collection

- Web Crawling
- Repository (e.g., Kaggle)
- Collection

02 Data Understanding

- Domain/data Understanding
- Statistical Analysis
- Feature Selection

03 Preprocessing

- Missing information, Filtering bad examples
- normalization
- Train/Test Preparation
  - Train data not in the Test data
  - Class distribution

04 Train

- Model Selection
- Meta parameter Tuning
- Enough training until performance converges

05 Test

- Evaluation Metrics
- Baseline

Q&A