SKC
텍스트 전처리

Natural Language Processing & AI Lab.,
Korea University

발표자: 허윤아, 박찬준

KOREA UNIVERSITY

Natural Language Processing & Artificial Intelligence

NLP Basics

Overview

자연언어란?

자연언어 (Natural language)

- 자연언어란?
  - 인간 고유의 언어
  - 정보전달의 수단
  - 인공언어에 대응되는 개념
  - 특정 집단에서 사용되는 모국어의 집합
    - 한국어, 영어, 불어, 독일어, 스페인어, 일본어, 중국어 등

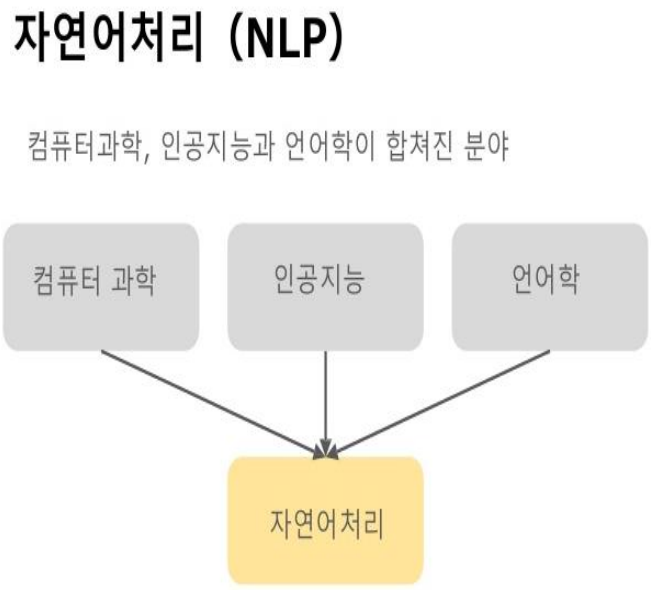
- 인공언어란?
  - 특정 목적을 위해 인위적으로 만든 언어
  - 자연언어에 비해 엄격한 구문을 가짐
  - 형식언어, 에스페란토어, 프로그래밍 언어

자연언어처리란?

자연언어처리

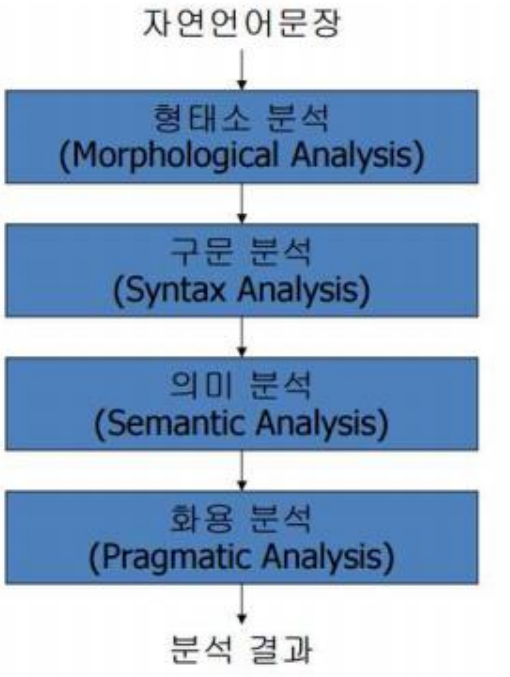
- 자연 언어 처리/ 자연어 처리
  - 컴퓨터를 통하여 인간의 언어를 처리하고 이용하려는 학문 분야
  - 인간의 언어를 이해하고, 이를 바탕으로 각종 정보처리에 적용함으로써 보다 빠르고 편리한 정보 획득

- 자연언어처리 응용 분야
  - 인간의 언어가 사용되는 실세계의 모든 영역
  - 정보검색, 질의응답 시스템
  - 기계번역, 자동통역
  - 문서작성, 문서요약, 분서분류, 철자오류 검색 및 수정, 문법 오류 검사



자연언어처리의 단계

자연언어처리의 단계



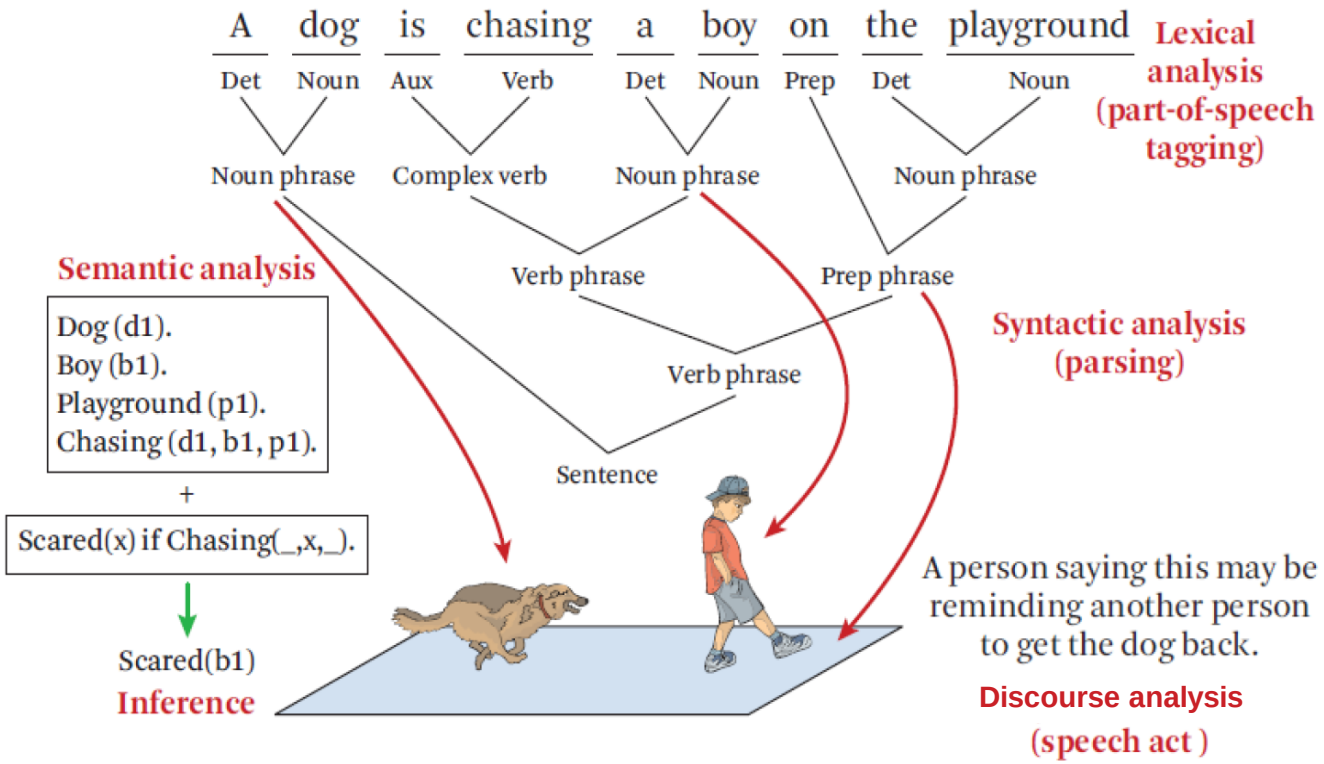
General steps in NLP

- Lexical Analysis
  - Word(lexicon), Morphology, word segmentation
- Syntax Analysis
  - Sentence structure, phrase, grammar,…
- Semantic Analysis
  - Meaning, execute commands
- Discourse Analysis
  - Meaning of a text
  - Relationship between sentences

    Ex) I disagree and so does John. (does-> disagree)

자연언어처리의 단계

- General steps in NLP



Lexical Analysis

- Lexical analysis의 필요성
  - 입력된 문장을 잘 분할해서 효율성을 높이기 위함

- Lexical analysis에서의 주요 요인
  - Sentence splitting : 마침표(.), 느낌표(!), 물음표(?) 등을 기준으로 분리
  - Tokenizing : 문서나 문장을 분석하기 좋도록 나눔 (띄어쓰기 또는 형태소 단위로…)
  - Morphological : 토큰들을 좀 더 일반적인 형태로 분석해 단어 수를 줄여 분석의 효율성을 높임 (가장 작은 의미단위로 토큰화 함)
  - Stemming (‘cars’,’car’=>’car’), lemmatization (단어를 원형으로),…

Syntax Analysis

Syntax analysis의 필요성

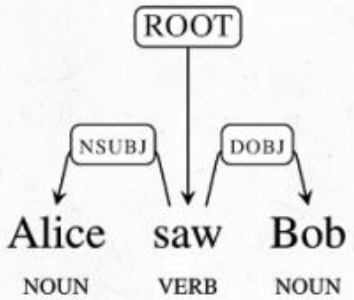
- 언어는 문장구성에 규칙이 필요함
- 문장 구성을 위한 규칙 / 문법을 구성

Syntax analysis (구문분석)

- 각각의 어절단위로 구분, 해당 tag 부여 (parsing tree이용)
- 의존성 구분분석 트리로 표현 (dependency parsing)
- Saw(root)를 기준으로 Alice(subject), Bob(object) 관계를 나타냄
- 이 문법적 관계들은 아주 직접적인 연관이 있음

POS	UMLS	Penn Tree Bank Tag	Example
Noun	noun	NN, NNS, NNP, NNPS	table
Adjective	adj	JJ, JJR, JJS	blue
Adverb	adv	RB, RBR, RBS, WRB, RP	quickly
Pronoun	pron	PRP, PRP$, WP, WP$	she
Verb	verb	VB, VBD, VBG, VBN, VBP, VBZ	wrote
Determiner	det	DT, PDT, WDT	the
Preposition	prep	IN	with
Conjunction	conj	CC	and
Auxiliary	aux	VB, VBD, VBG, VBN, VBP, VBZ	does
Modal	modal	MD	could
Complement	compl	IN	that

Part-of-speech tagging



Dependency parsing

Semantic Analysis

Semantic analysis의 필요성

- 규칙에 따라 문장은 만들었는데 문장이 의미적으로 올바른 것인지 알아야 함
- 사람이 사과를 먹는다. (O)
- 사람이 비행기를 먹는다. (X)

Semantic analysis란?

- Syntactic + meaning
- Two levels (lexical semantics)
- Representing meaning of words
- Word sense disambiguation (word bank)

- Compositional semantics
- How words combined to form a larger meaning

"6시에 KBS에서 뭐 하니?"

Semantic analysis

Question focus : 프로그램
Channel : KBS
Begin_time : 18:00

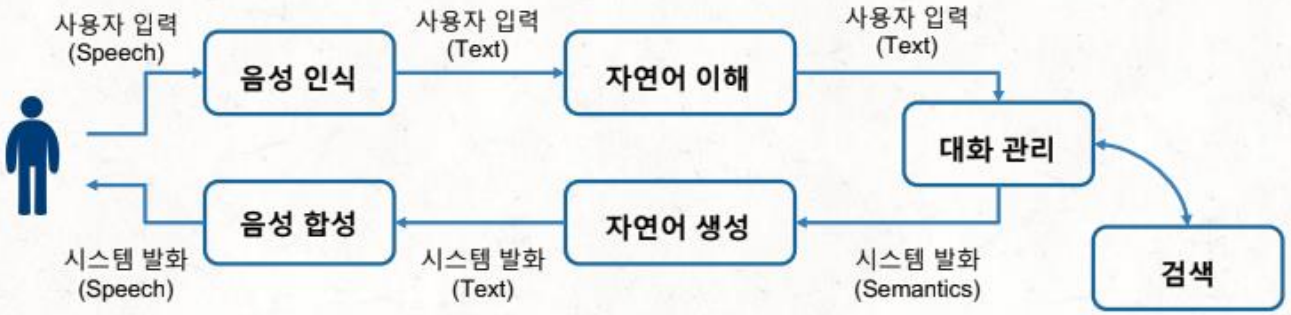
Discourse Analysis

Discourse analysis의 필요성

- 대화의 흐름을 파악하여 발화자의 의도에 맞도록 응답해야 함

Discourse analysis란?

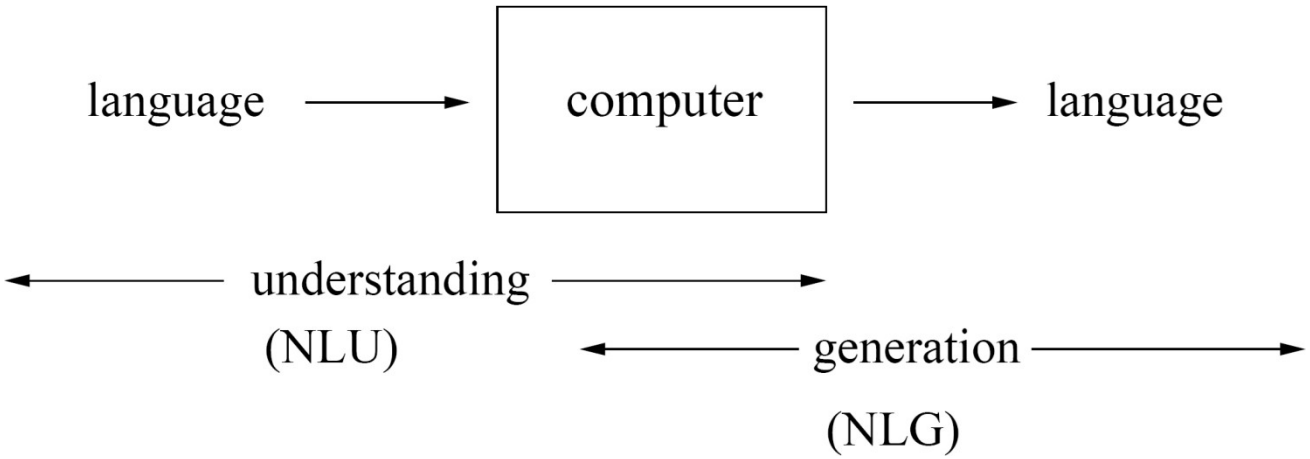
- 대화의 흐름상 어떤 의미를 가지는지를 찾음
- 문맥구조 분석 (문장들의 연관 관계)
- 의도분석 (전후관계를 통한 실제 의도)
- 대화 분석 (대표적인 담화분석)



대화 시스템 프로세스

자연언어처리의 단계

- One simple (but practical) answer
  - Computer using natural language as input and/or output



  - Or components enabling such a computer

자연언어처리를 위한 언어학

Overview

음운론

Phonetics (음성학) & Phonology (음운론)

- The study of language sounds, how they are physically formed and systems of discrete sounds
  - disconnect => dis-k&-'nekt
  - "It is easy to recognize speech."
  - "It is easy to wreck a nice beach."
- 음성인식
  - Signal to symbol

음소

더 이상 작게 나눌 수 없는 음운론상의 최소 단위

형태론

어절, 단어, 형태소

- 어절
  양쪽에 공백을 갖는 띄어쓰기 단위의 문자열
- 단어 / 형태소
  단일 품사를 갖는 단위 / 사전에 등록되어 있는 색인어의 집합

예: 나는 책을 읽었다.
파릇파릇한 싹이 나는 계절이다.
하늘을 나는 새를 보라.
I tried to go to school.
He tries to pass the exam.

나 + 는
날다 + 는
나다 + 는

형태소
의미를 가지는 언어 단위 중 가장 작은 단위

의미 혹은 문법적 기능의 최소단위

형태소 분석

형태소 분석

- 형태소 분석이란?
  - 형태소를 비롯하여, 어근, 접두사/접미사, 품사(POS, Part-of-Speech) 등 다양한 언어적 속성의 구조를 파악하는 것
- Mecab, Twiiter, Komoran, Hannanum, KoNLPy 등

```
pprint(kkma.pos(a))
```

```
문장을 입력하세요:세종대왕님은 글을 만드셨습니다.
[('세종', 'NNG'),
 ('대왕', 'NNG'),
 ('님', 'XSN'),
 ('은', 'JX'),
 ('글', 'NNG'),
 ('을', 'JKO'),
 ('만들', 'VV'),
 ('시', 'EPH'),
 ('었', 'EPT'),
 ('습니다', 'EFN'),
 ('.', 'SF')]
```

형태소 분석

형태소 분석 (Morphological Analysis)

- 입력된 문자열을 분석하여
  형태소(morpheme)라는 최소 의미 단위로 분리
- 사전 정보와 형태소 결합 정보 이용
- 정규 문법(Regular Grammar)으로 분석 가능
- 언어에 따라 난이도가 다름
  - 영어, 불어 : 쉬움
  - 한국어, 일본어, 아랍어, 터키어 : 어려움

형태소 분석의 어려움

형태론적 다양성

- 첨가어
  - 한국어, 일본어, 터키어 등
  - 다수의 형태소가 결합하여 어절 형성
  - 터키어는 평균 7개의 형태소가 결합
- 굴절어
  - 라틴어 (영어, 불어 등은 첨가어와 굴절어의 특징이 모두 있음)
  - 어간이 변함 (영어의 예 : run, ran, run)
- 스와히리어
  - 수(number)를 위한 형태소가 문두에 붙음
  - (예) 사람 : m+tu (단수), wa+tu (복수)
    나무 : m+ti (단수), mi+ti (복수)
- 아랍어
  - 자음이 어간이고 모음이 시제, 수 등을 표현
  - (예) ktb(쓰다) kAtAb(능동) KUtIb(수동)
    kttb(쓰게하다) kAttAb(능동)KUttIb(수동)

통사적 다양성

- Postfix 언어 (Head-Final Languages)
  - 동사가 문장의 뒤에 위치
  - 한국어, 일본어 등
- Infix 언어
  - 동사가 문장의 중간에 위치
  - 영어, 불어 등
- Prefix 언어
  - 동사가 문장의 처음에 위치
  - 아일랜드어

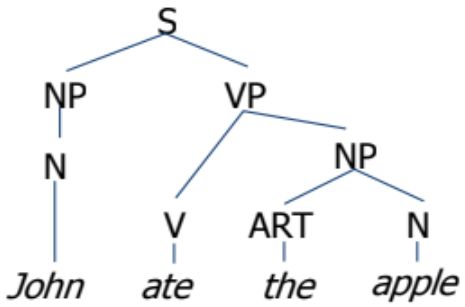
형태소 분석의 어려움

형태소 분석의 난점

- 중의성 (ambiguity)
  - "감기는"의 분석 결과
    - 감기(명사:cold) + 는(조사)
    - 감(동사 어간) + 기(명사화 어미) + 는 (조사)
    - 감(동사 어간) + 기는(어미)
- 접두사, 접미사 처리
- 고유명사, 사전에 등록되지 않은 단어 처리
  - 한국어, 독일어처럼 복합명사 내의 명사를 띄우지 않거나, 일본어처럼 띄어쓰기가 없으면 더욱 어려워짐
- 한국어 형태소 결합의 예 ("친구에게서였었다라고")
  - 친구(명사) + 에게(조사) + 서(조사) + 이(서술격조사) + 었(과거시제어미) + 었(회상어미) + 다(어말어미) + 라고(인용격조사)

문법, 구문 분석

- 문법 (Grammar) :
  - 문장의 구조적 성질을 규칙으로 표현한 것
- 구문 분석기 (Parser) :
  - 문법을 이용하여 문장의 구조를 찾아내는 process
  - 문장의 구문 구조는 Tree 형태로 표현할 수 있다. 즉, 몇 개의 형태소들이 모여서 구문 요소(구: phrase)를 이루고, 그 구문 요소들간의 결합구조를 Tree형태로써 구문 구조를 이루게 된다.



문법 (Grammars)

- Grammar : a set of rewrite rules

(ex)
S → NP VP
NP → ART N
NP → N
VP → V NP

- Context Free Grammar (CFG) :
  - 각 rule의 LHS(Left-Hand side)가 하나의 symbol로 이루어진 문법 규칙
- Grammar Rule 을 이용해서 문장(sentence)을 생성할 수도 있고(sentence generation), 분석할 수도 있다(sentence parsing).

구문론

Sentence Generation

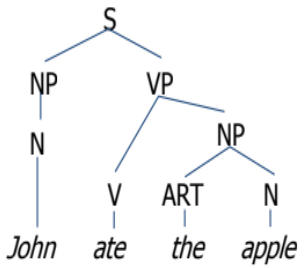
(ex) By rewrite rule

S → NP VP
→ N VP
→ *John* VP
→ *John* V NP
→ *John* *ate* ART N
→ *John* *ate* the N
→ *John* *ate* *the* *apple.*

Bottom-up Parsing

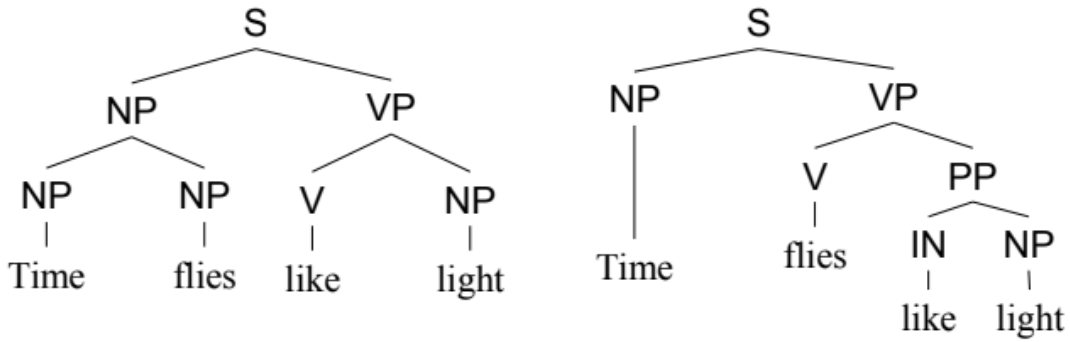
(ex) *John ate the apple.*

→ N V ART N
→ NP V ART N
→ NP V NP
→ NP VP
→ S



구문 분석의 어려움

구문 분석 - Structural Ambiguities



- Structural Ambiguities
  - Time flies like light. ⇒ 2가지 이상의 구조로 분석됨
    - flies (noun *or* verb), like(verb *or* preposition)
  - A man see a woman with a telescope on the hill. ⇒ 5가지 이상

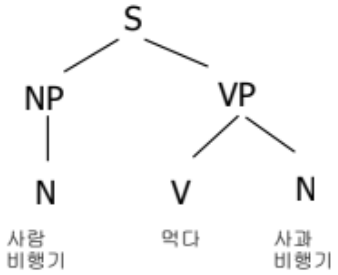
의미 분석 (Semantic Analysis)

- 통사 분석 결과에 해석을 가하여 문장이 가진 의미를 분석
- 형태소가 가진 의미를 표현하는 지식 표현 기법이 요구됨
- 통사적으로 옳으나 의미적으로 틀린 문장이 있을 수 있음
  - 돌이 걸어간다 (*cf*. 사람이 걸어간다)
  - 바람이 달린다 (*cf*. 말이 달린다)
- Ambiguity
  - 말이 많다 (horse, speech)

의미 분석 – cont'd

- 문법적으로는 맞지만 의미적으로 틀린 문장들
  - 사람이 사과를 먹는다. (o)
  - 사람이 비행기를 먹는다. (x)
  - 비행기가 사과를 먹는다. (x)

구문 구조



의미적 제약

[먹다
[agent : 먹을수 있는 주체
object : 먹을 수 있는 대상
....]]

화용 분석 (Pragmatic Analysis)

- 문장이 실세계(real world)와 가지는 연관관계 분석
- 실세계 지식과 상식의 표현이 요구됨
- 지시(anaphora), 간접화법(indirect speech act) 등의 분석
  - Anaphora : 대명사의 지시 대상

    The city councilmen refused the women a permit because
    (1) *they* feared violence.
    (2) *they* advocated revolution.

  - Speech Act : 상대방에게 행동을 요구하는 언어 행위

    Can you give me a salt?

    Would you mind opening the window?

감사합니다.

박찬준
bcj1210@naver.com