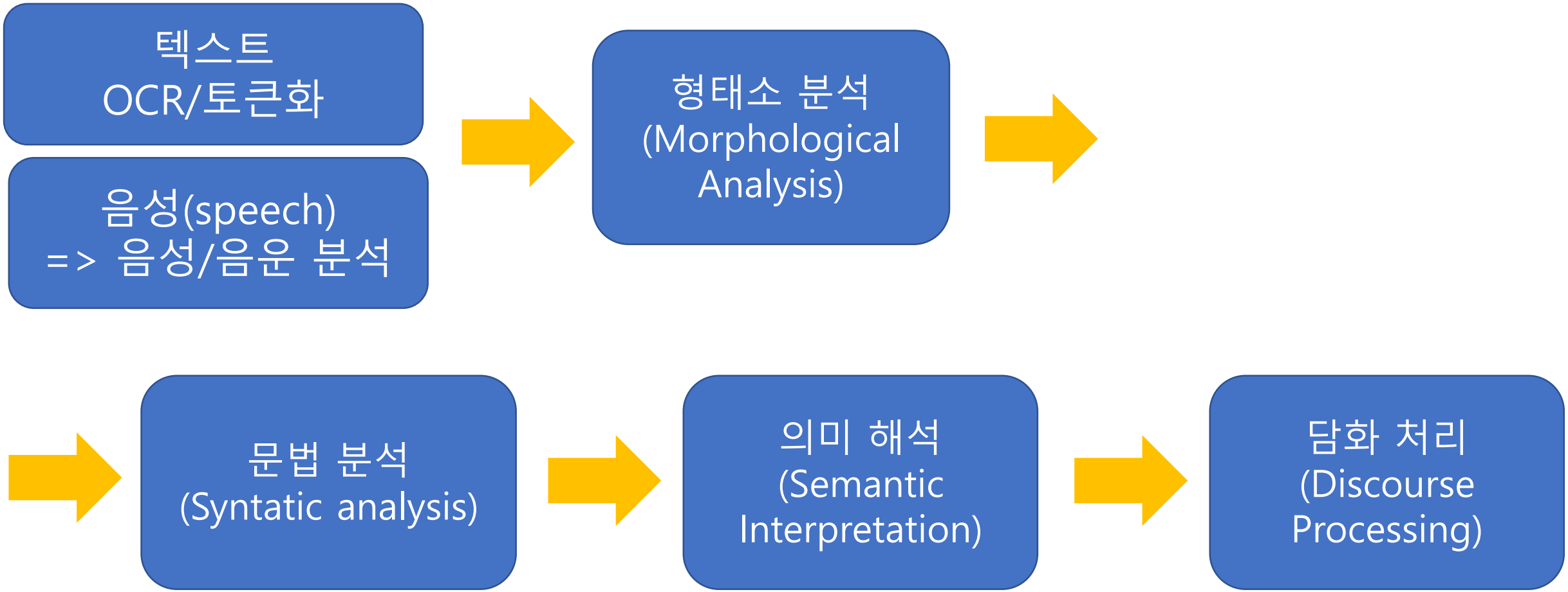
자연어 처리
(with Python)

1-1 자연어 처리 단계


텍스트
OCR/토큰화
음성(speech)
=> 음성/음운 분석
형태소 분석
(Morphological
Analysis)
문법 분석
(Syntatic analysis)
의미 해석
(Semantic
Interpretation)
담화 처리
(Discourse
Processing)

1-2 자연어 처리의 응용분야

(1) 간단한 응용

A. 철자 검사

B. 키워드 검색

C. 유사어 검색

(2) 중급 응용

A. 웹 사이트로부터 정보 추출

B. 감정 분석

(3) 고급 응용

A. 기계 번역

B. 음성 인식

C. 챗봇/음성 대화 시스템

D. 복잡한 질의 응답

1-3 형태소에 대해 알아보기

(1) 형태소는 의미를 전달하는 기본 단위이다.

(2) 종류

A. 어간

B. 접사

C. 영문

1-4 기본 용어 정리

(1) 음소(Phoneme)

A. 더 이상 작게 나눌 수 없는 음운론상의 최소 단위 (예) ㅎ ㅏ ㄱ

B. 하나 이상의 음소가 모여서 음절을 이룬다.

(2) 음절(Syllable)

A. 몇 개의 음소로 이루어지며, 모음은 단독으로 한 음절이 되기도 함.

B. 하나의 종합된 음의 느낌을 주는 것 (예) 합

(3) 형태소(Morpheme)

A. 뜻을 가진 가장 작은 말의 단위

B. ‘이야기 책’의 ‘이야기‘, ‘책 '

(4) 단어(Word)

A. 형태소들로 구성되는 분리하여 자립적으로 쓸 수 있는 말.

1-4 기본 용어 정리

(5) 품사(Part-of-Speech, POS, Word class)

A. 단어를 기능, 형태, 의미에 따라 나눈 갈래

B. 9가지 또는 22가지로 분류

(6) 품사 태깅(Part of Speech Tagging)

A. 문장을 형태소나 품사별로 나누어 줌.

(7) 문장(통사, statement)

A. 생각이나 감정을 말과 글로 표현할 때 완결된 내용을 나타내는 최소 단위

(8) 담화(discourse)

A. 둘 이상의 문장이 연속되어 이루어지는 말의 단위

(9) 말뭉치(코퍼스, Corpus)

A. 대량의 구조화된 텍스트(a large and structured set of texts)

1-5 자연어 처리 라이브러리

(1) NLTK

A. 자연어 처리를 위한 오픈소스

(2) KoNLPy

A. 한국어 자연어 처리를 위한 파이썬 패키지

(3) mecab

A. 자연어 처리 C++ 라이브러리

B. 띄워쓰기 처리가 없음.

C. 은전 한 잎.

D. KoNLP에서 은전 한 잎을 파이썬으로 옮김.

(4) FudanNLP

A. 중국어 처리 라이브러리

B. Github : https://github.com/xpqiu/fnlp

1-5 자연어 처리 라이브러리

(5) Standford Natural Language Processing Group NLP

A. http://nlp.standard.edu/software/index.html

1-6 자연어 처리 라이브러리(NLTK)

(1) 구성 요소

A. 50개 이상의 말뭉치(corpus)와 어휘(lexicon)가 있음.

B. 분류, 토큰화, 스테밍(Stemming), 태깅(tagging), 구문 분석(parsing), 의미 추리(semantic reasoning) 라이브러리 API

C. NLTK 공식 추천 온라인 Book

http://www.nltk.org/book (파이썬 3.x용)

http://www.nltk.org/book_1ed (파이썬 2.x용)

1-6 자연어 처리 라이브러리(NLTK)

(1) 패키지 구성 요소

데이터

A. nltk.data : 말뭉치 문법, 기타 저장된 객체와 같은 NLTK 지원 파일의 로드
B. nltk.downloader : 말뭉치, 모델, 기타 패키지 다운로드
C. nltk.corpus : 말뭉치(Corpus)와 어휘(lexicon)에 대한 표준화된 API

텍스트 분석

A. nltk.probability : 확률 정보의 표현과 처리를 위한 클래스들
B. nltk.tokenize : 토큰화
C. nltk.stem : 스테밍
D. nltk.collocations : 연어(collocation)검출: t-test, chi-squared, point-wise mutual information

1-6 자연어 처리 라이브러리(NLTK)

(1) 패키지 구성 요소

데이터

A. nltk.data : 말뭉치 문법, 기타 저장된 객체와 같은 NLTK 지원 파일의 로드

B. nltk.downloader : 말뭉치, 모델, 기타 패키지 다운로드

C. nltk.corpus : 말뭉치(Corpus)와 어휘(lexicon)에 대한 표준화된 API

텍스트 분석

A. nltk.probability : 확률 정보의 표현과 처리를 위한 클래스들

B. nltk.tokenize : 토큰화

C. nltk.stem : 스테밍

D. nltk.collocations : 연어(collocation)검출: t-test, chi-squared, point-wise mutual information

연어? 어떤 언어 내에서 특정한 뜻을 나타낼 때 흔히 함께 쓰이는 단어들의 결합.

1-6 자연어 처리 라이브러리(NLTK)

(1) 패키지 구성 요소

텍스트 분석

A. nltk.tag : 품사(Part-of-speech:POS) 태깅(tagging) : n-gram, backoff, Brill, HMM, Tnt

B. nltk.chunk : 정규 표현, n-gram, 개체명(고유명사)

구문 분석

nltk.grammar, nltk.parse, nltk.app, nltk.ccg

의미 분석

Nltk.sem, nltk.inference

기계학습

nltk.classify : 결정 트리, 나이브 베이즈, maximum entropy

1-6 자연어 처리 라이브러리(NLTK)

(1) 패키지 구성 요소

기계학습

A. nltk.classify 패키지 : 결정 트리, 나이브 베이즈, maximum entropy

B. nltk.cluster 패키지 : [군집] k-means, EM

C. nltk.tbl 패키지 : [변환](transformation) 기반 학습

D. nltk.metrics 패키지 : precision, recall, agreement, coefficients

자연어 처리 응용

A. nltk.sentiment 패키지 : 감성 분석

B. nltk.translate 패키지 : 기계번역

C. nltk.chat 패키지 : 간단한 챗봇(chatbot) API

1-7 토큰(Token)

(1) 토큰(token) : 의미를 갖는 문자열(단어, 절, 문장 등)

(2) 토크나이징(tokenizing) : 토큰을 나누는 작업

(3) 영문-공백

(4) 한글 – 합성어, 조사합성

(5) 작업기준

1-8 정규식 참조

(1) http://www.nextree.co.kr/p4327

2-1 알아보기

(1) Stemming

(2) Tagging

(3) WordCloud

(4) StopWord

(5) Tf-idf

2-1 알아보기

(1) Stemming : 정규화

(2) Tagging : 토큰에 대한 품사 태깅.

(3) WordCloud

(4) StopWord

(5) Tf-idf

2-3 형태소 분석기의 성능 비교

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/05/10/postag/

2-4 KoNlpy 사용

1. KKMA
2. Hannanum
3. Twitter : OpenKoreanText 오픈 소스 한국어 처리기
4. Eunjeon : 은전한닢 프로젝트
5. KOMORAN : Junsoo Shin 님의 코모란
6. 트위터 : 빠른 분석
7. 꼬꼬마 : 정확한 품사 정보 필요
8. 정확성, 시간 모두 중요 : 코모란

2-5 한국어 형태소 분석기

C/C++

- KTS (1995) GPL v2
  - 이상호, 서정연, 오영환 (KAIST & 서강대)
  - code
- MACH (2002) custom
  - 심광섭 (성신여대)
- MeCab-ko (2013) GPL LGPL BSD
  - 이용운, 유영호

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-5 한국어 형태소 분석기

자바

- 아리랑 (2009) Apache v2
  - 이수명
  - code
- 한나눔 (1999) GPL v3
  - KAIST 최기선 교수 연구팀
  - code, docs
- 꼬꼬마 (2010) GPL v2
  - 서울대 이상구 교수 연구팀
  - 동적 프로그래밍을 이용해 형태소 후보를 찾음
  - 형태소의 주변을 확인하고, 몇몇 휴리스틱을 사용하고, HMM을 사용하는 방식으로 품사를 태깅함
  - 개발자 블로그: 이동주
- KOMORAN (2013) Apache v2
  - By *shineware*

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-5 한국어 형태소 분석기

파이썬

- KoNLPy (2014) GPL v3
  - 박은정 (서울대)
- UMorpheme (2014) MIT
  - 김경훈 (UNIST)

R

- KoNLP (2011) GPL v3
  - 전희원

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-5 한국어 형태소 분석기

그 외

- K-LIWC (아주대)
- KRISTAL-IRMS (KISTI)
  - 개발 후기
- Korean XTAG (UPenn)
- HAM (국민대)
- POSTAG/K (포스텍)
- Speller (부산대)
- UTagger (울산대)
- (No name) (고려대)

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-5 한국어 형태소 분석기

다른 NLP 도구

- Hangulize - By Heungsub Lee Python
  - Hangul transcription tool to 38+ languages
- Hanja - By Sumin Byeon Python
  - Hanja to hangul transcriptor
- Jamo - By Joshua Dong Python
  - Hangul syllable decomposition and synthesis
- KoreanParser - By DongHyun Choi, Jungyeul Park, Key-Sun Choi (KAIST) Java
  - 언어 파서
- Korean - By Heungsub Lee Python
  - Package for attaching particles (josa) in sentences

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-5 한국어 형태소 분석기

말뭉치

- 연세 말뭉치, 연세대, 1987.
  - 1960년 이후 한국어에 대한 4200만 어절
- 고려대학교 한국어 말뭉치, 1995
  - 1970-90년대 한국어에 대한 1000만 어절
- HANTEC 2.0, KISTI & 충남대, 1998-2003.
  - 12만 개의 테스트 문서 (237MB)
  - QA를 위한 50개의 TREC 형태 질의
- HKIB-40075, KISTI & 한국일보, 2002.
  - 텍스트 분류를 위한 40,075 테스트 문서 (88MB)
- KAIST Corpus, KAIST, 1997-2005.
- Sejong Corpus, National Institute of the Korean Language, 1998-2007.

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-6 NLP 관련 사이트

- Google NLP publications
- Lingpipe
- Microsoft NLP group (Redmond)
- 부산대 NLP 관련사이트 목록
- Sejong semantic search system
- 한글 및 한국어 정보처리 학술대회

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조

2-6 NLP 관련 사이트

- Google NLP publications
- Lingpipe
- Microsoft NLP group (Redmond)
- 부산대 NLP 관련사이트 목록
- Sejong semantic search system
- 한글 및 한국어 정보처리 학술대회

https://konlpy-ko.readthedocs.io/ko/v0.4.3/references/#corpora 참조