2025-1 한국어 정보 처리

# 데이터 수집 및 전처리 1

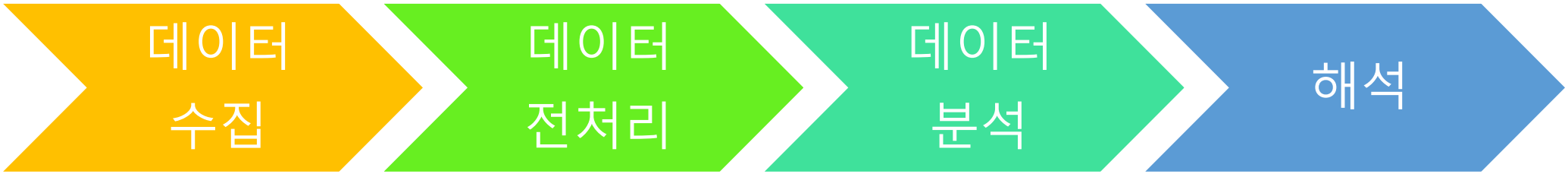
전태희

taeheejeon22@gmail.com

# 목차

# 1. 데이터 기반 연구의 시작: 자연어 처리란?

# 데이터를 활용하는 연구의 흐름




- 언어 데이터를 활용하는 예
  - 뉴스 기사 수집 -> 전처리(토큰화, 오류 수정 등) -> 토픽 모델링 -> 사회적 관심사 해석
- 두 가지 방향
  - 학문적 접근: 언어 현상 설명, 이론 검증
    - 시대별 언어 변화 추적, 방언 비교 분석, 담화 구조 분석, 한국어 학습자 오류 분석
  - 자연어 처리: 자동화, 정보 추출, AI 응용 등
    - 키워드 추출, 문서 분류, 감성 분석, 자동 요약, 기계 번역, 음성 인식/합성 등

# 자연어 (Natural Language)

- 의도적인 계획이 관여하지 않으면서, 사용, 반복, 변화의 과정을 거쳐 인간 사회에서 자연스럽게 발생하는 언어(Natural language - Wikipedia)
  - 자연어가 아닌 것
    - 프로그래밍 언어와 같은 인공어
    - 에스페란토(Esperanto) 등의 인공의 국제어
    - 고래, 벌 등 인간이 아닌 존재가 사용하는 의사소통 체계

✓ 즉 인공적인 것이 아님 and 인간이 씀

# 자연어 처리 (Natural Language Processing)

- 컴퓨터 과학과 정보 검색의 학제적 하위 분야로, 컴퓨터가 인간의 언어를 다룰 수 있는 능력을 갖추도록 하는 것과 주로 관련됨 (Natural language processing - Wikipedia)
- 기계 학습을 활용하여, 컴퓨터가 인간의 언어를 이해하고 인간의 언어로 소통할 수 있도록 하는 컴퓨터 과학 및 인공 지능의 하위 분야 (What Is NLP (Natural Language Processing)? | IBM)

# 기본 절차 [1/3]

1. 데이터 수집
   - ○다양한 소스로부터 데이터를 수집하는 단계
   - ○코퍼스 구축으로 볼 수 있음
2. 전처리(Pre-processing)
   - ○가공되지 않은 원데이터를 가공하는 단계
   - ○데이터에 존재하는 오류를 제거하고, 데이터를 표준화하여 컴퓨터가 텍스트를 처리하기 쉽도록 함
   - ○영어에 대해서는 모든 문자열을 소문자화(lowering)하고, 실질적 의미 없이 형식적으로 쓰이는 단어들(Stop words. 관사, be 동사 등)을 제거하는 등의 방법들이 이용되어 옴

# 기본 절차 [2/3]

## 3. 토큰화(Tokenization)

- 연속적인 문자열을 컴퓨터가 의미 있는 세부 단위인 토큰으로 분절하는 작업
- 문장, 단어, 형태소 등으로 토큰화함
- 전처리의 일부로 취급되기도 함

## 4. 텍스트 표상 (Text Representation)

- 텍스트를 수치 데이터로 변환하여 기계가 처리할 수 있도록 하는 작업
- Word2Vec, BERT, GPT 등에서 이용되는 임베딩(Embedding)이 이에 해당함

# 기본 절차 [3/3]

5. 모델 구축
   - ○데이터를 이용해 모델 학습
   - ○감성 분석, 스팸 탐지 등의 특정 태스크만을 위한 모델을 구축할 수 있으며 최근의 LLM과 같이 일반적인 목적의 모델을 구축할 수도 있음
   - ○LLM에서는 텍스트 표상이 본 단계에 포함됨
6. 평가
   - ○모델의 성능을 태스크에 따라 정확도(Accuracy), F1-Score, BLEU 등의 척도(Metric)로 평가함
   - ○LLM의 성능을 평가하는 것은 다소 복잡함

# 2. 코퍼스에 대한 이해

# 코퍼스란?

- 디지털 환경에서 생성되었거나 기존 문헌을 디지털화하여 구성된 데이터 세트 (Wikipedia, Text corpus)
- 발화된 언어 자료의 집합체로서, 1) 어떤 언어 연구의 목적에 부합하며 2) 균형성과 대표성을 가진 3) 전산화된 자료 (강범모, 2008, 언어 기술을 위한 코퍼스의 구축과 빈도(통계) 활용)

- 협의의 코퍼스
  - 언어 연구를 목적으로 균형성과 대표성을 고려해 정교하게 설계, 구축됨
  - 언어 분석 정보: 원문(raw text), 형태 분석, 구문 분석, 의미 분석, 구문 분석
- 광의의 코퍼스
  - 다양한 목적으로 활용 가능한 전산화된 언어 자료로 균형성과 대표성이 반드시 고려되는 것은 아님
  - 블로그, 댓글, 뉴스 기사, SNS 텍스트 등

# 코퍼스의 필요성

- 언어의 빈도나 패턴에 대한 인간의 직관은 불완전함

제4절 단수 표준어

제17항 비슷한 발음의 몇 형태가 쓰일 경우, 그 의미에 아무런 차이가 없고, 그중 하나가 더 널리 쓰이면, 그 한 형태만을 표준어로 삼는다.(ㄱ을 표준어로 삼고, ㄴ을 버림.)

| ㄱ | ㄴ | 비고 |
|---|---|---|
| 거든-그리다 | 거둥-그리다 | 1. 거든하게 거두어 싸다.<br>2. 작은말은 '가든-그리다'임. |
| 구어-박다 | 구워-박다 | 사람이 한 군데에서만 지내다. |
| 귀-고리 | 귀엣-고리 | |
| 귀-띔 | 귀-틤 | |
| 귀-지 | 귀에-지 | |

- 객관적 언어 분석을 위한 도구
  - 언어 현상을 실제 데이터에 기반해 분석함으로써 통계 기반의 정량적 연구가 가능
    - Measure what is measurable, and make measurable what is not so.

Galileo Galilei (1564-1642)

# 코퍼스의 균형성과 대표성

- 균형성(Balance)
  - 다양한 장르, 매체, 시기, 지역, 화자의 연령/성별 등을 고려해 균형 있게 구성
  - 가령 특정 장르에 편향되면 그 장르의 특성이 과도하게 반영될 수 있음
    - 예: 신문 기사 10%, 소설 10%, 블로그 30%, SNS 50%로 구성된 코퍼스
      - ➢ 비격식적, 구어적 특성이 두드러지게 됨
- 대표성(Representativeness)
  - 언어 사용의 모집단(population)
  - 언어 사용의 표본(sample)으로서의 코퍼스
    - 선거 여론 조사에서 표본이 인구 전체를 대변할 수 있듯, 코퍼스도 마찬가지
  - ✓모든 코퍼스는 하나의 샘플로, 샘플링이 잘되어야 전체 언어의 모습을 잘 관찰할 수 있다!

# 코퍼스의 유형

- 텍스트 성격
  - 문어 코퍼스: 신문, 소설, 학술지 등
  - 구어 코퍼스: 일상 대화, 인터뷰, 방송 대본 등
- 구성 목적
  - 일반 목적: 다양한 장르, 주제를 포함하며 균형성, 대표형이 중요
  - 특수 목적: 의료, 법률 등 특정 주제, 장르 중심
- 언어 분석 정보
  - 원시 코퍼스 (raw corpus): 텍스트에 부가 정보를 덧붙이지 않은 코퍼스
  - 주석 코퍼스 (Annotated corpus): 텍스트에 형태, 구문, 의미 등 언어학적 주석을 덧붙여 언어 분석에 용이하도록 만든 코퍼스

# 코퍼스에 대한 말들

- I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. (Fillmore, 1992, 'Corpus linguistics' or 'Computer-aided Armchair Linguistics')
  - 코퍼스의 한계
  - 코퍼스의 필요성
- Chomsky
  - "I live in Dayton, Ohio."보다 "I live in New York."이라는 문장이 코퍼스에 훨씬 많이 나타날 것인데, 이것이 언어 기술에 무슨 의미가 있겠는가?

✓직관 중심의 언어학과 데이터 기반 언어학 사이의 균형이 필요

# 3. 대표 코퍼스 및 연구 사례

# 대표 코퍼스: Brown corpus [1/2]

- 최초의 전산 코퍼스 (1961, Brown University)
- 약 1백만 단어 규모
  - 균형성과 대표성을 고려해 15개 장르로 분류된 500개 텍스트
    - 텍스트별 2,000개 단어 추출
- 문어 미국 영어의 사용 특성을 반영함
- 품사(Part-of-speech) 주석이 갖추어짐

# 대표 코퍼스: Brown corpus [2/2]

- 구성
  - 신문: 18%
  - 책-정보: 57%
  - 책-문학: 25%
- 층화 표본 추출 (Stratified Sampling)
  - 단순 무작위 추출 시, 특정 장르가 과표집되거나 누락되어 편향이 발생할 수 있음
  - 장르별 출판된 문서의 비율 파악하여 장르별 텍스트 후보를 선정 후, 장르별 후보 중  무작위로 표본 텍스트 선정
  - 결과적으로 모든 장르의 텍스트가 현실의 데이터 비율에 맞게 추출됨
- 의의
  - 코퍼스 설계의 표준을 제시
  - 빈도 기반 언어 분석을 위한 자료로서 활용되어, 언어 교육, 언어 심리학 등에서 코퍼스의 효용 입증

# 대표 코퍼스: 세종 말뭉치 [1/4]

- 정부 주도로 구축된 대규모 한국어 코퍼스
- 구축기간
  - 1998~2007년 (10년)
- 구축기관
  - 고려대학교 민족문화연구원
  - 연세대학교 언어정보개발연구원
- 주관 및 후원 기관
  - 문화관광부
  - 국립국어원
- 투입 예산
  - 코퍼스 구축 관련 분야 (도구 개발비 포함)
    - 약  57억 (세종 계획 전체 예산은 약 150억)
    - 구문 분석 코퍼스 구축 예산 1.5~2억 (추산)

# 대표 코퍼스: 세종 말뭉치 [2/4]

- 세종 말뭉치의 규모
  - 약 2억 어절
- 문어 코퍼스 위주
- 원시 코퍼스뿐 아니라
  형태, 의미, 구문 분석 코퍼스 자

(4) 세종말뭉치(21세기 세종계획)

| 말뭉치(코퍼스) | | 부문 | 어절 수 (단위: 백만) |
|---|---|---|---|
| 현대 문어 | | 원시 | 62.0 |
| | | 형태분석 | 15.0 |
| | | 형태의미분석 | 12.5 |
| | | 구문분석 | 0.8 |
| 현대 구어 | | 원시 | 3.7 |
| | | 형태분석 | 1.0 |
| 북한 및 해외 | | 원시 | 9.5 |
| | | 형태분석 | 1.6 |
| 역사 자료 | | 원시 | 5.6 |
| | | 형태분석 | 0.9 |
| 병렬 | 한·영 | 원시 | 4.8 |
| | | 형태분석 | 1.0 |
| | 한·일 | 원시 | 1.1 |
| | | 형태분석 | 0.3 |
| 현대 문어 보조 자료 | | 원시 | 75.0 |

# 대표 코퍼스: 세종 말뭉치 [3/4]

- 품사 태그 세트
  - 체언: NNG, NNP, NNB, NP, NR, XR
  - 용언: VV, VA, VX, VCP, VCN
  - 독립언: IC
  - 수식언: MAG, MAJ, MM
  - 접사: XPN, XSN, XSV, XSA
  - 조사: JKS, JKC, JKG, JKO, JKB, JKV, JKQ, JC, JX
  - 어미: EC, EF / ETN, ETM / EP
  - 기호: SF, SN, SL SP, SS 등
- 이 태그 세트는 많은 형태소 분석기에서 채택하고 있음
  - 형태소 분석기 Kkma, MeCab-ko, Khaii 등의 품사 태그 비교
    - https://lswkim322.gitbook.io/til/til-ml/boostcamp/p-stage-nlp/3-1

# 대표 코퍼스: 세종 말뭉치 [4/4]

▪입수 방법

○이전에는 인터넷을 통해 쉽게 다운로드할 수 있었으나 현재는 불가능함

○다음 링크에서 DVD 신청을 해야 함

- https://kli.korean.go.kr/corpus/boards/noticeView.do?page=0&recordId=417&boardId=&base.condition=board.title&base.keyword=%EC%84%B8%EC%A2%85&size=10

# 코퍼스 연구의 예 [1/3]

- 전지은(2022), 신문 코퍼스의 어휘 난이도 분석 및 활용 연구
  - ○한국어 학습자의 숙달도에 맞는 신문 텍스트를 활용하는 한국어 어휘 교육 방안 제시
  - ○물결 21 코퍼스의 웹 기반 코퍼스 분석 도구 활용

먼저 14번 텍스트는 유의어 '참여, 참석, 참가'가 쓰인 용례를 포함한 원본 텍스트 중 어휘 난이도가 가장 낮은 16.8로 중급(16.9) 수준인 학습자를 위한 자료로 활용할 수 있다.[11]

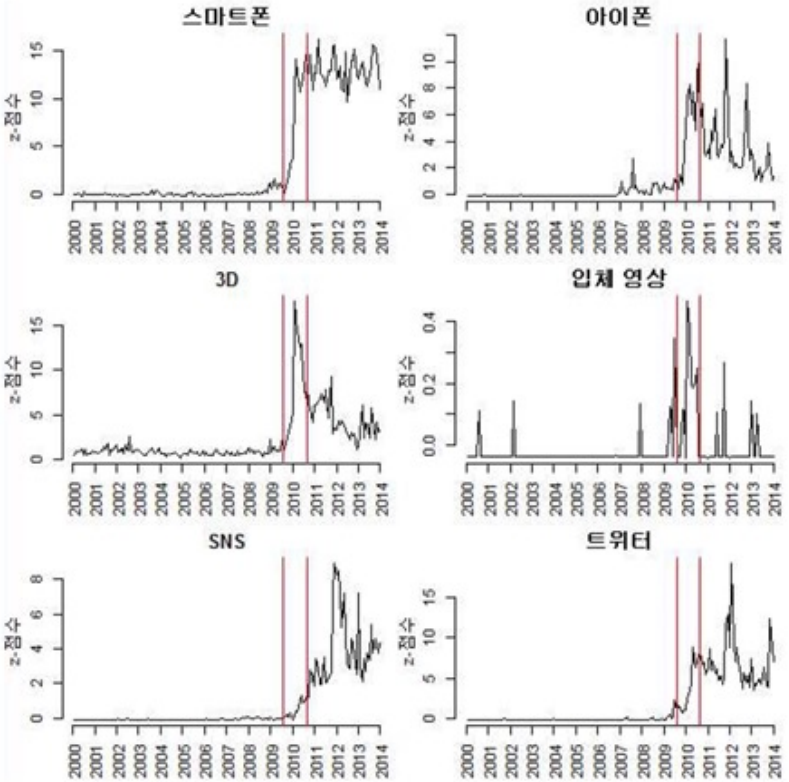
14번 텍스트: 어휘 난이도 16.8

"최소한(D)의 규제(C)없어(A) 우리 요구(B)와 정반대(B)" <H20080616_027>

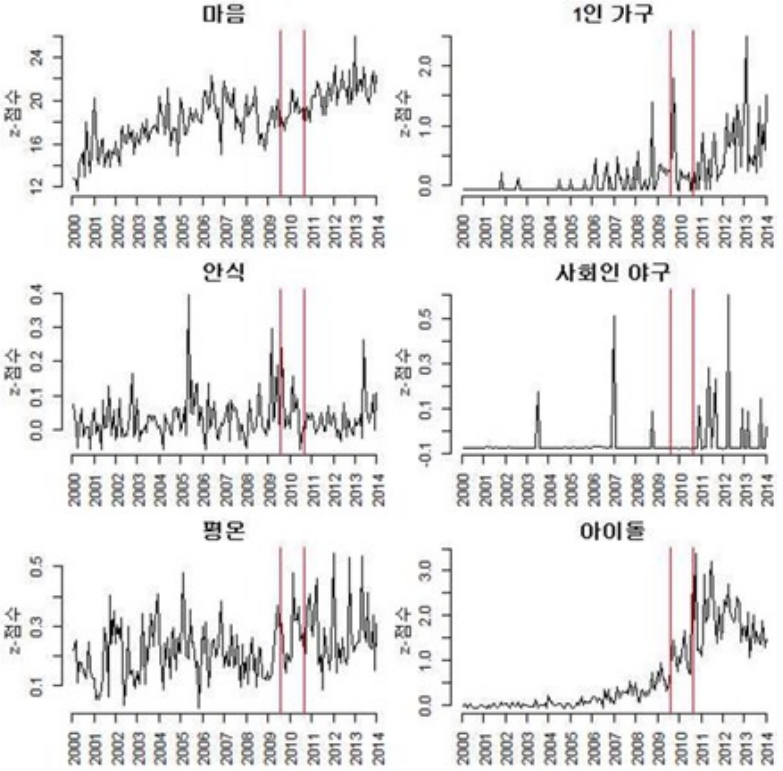
지난(A) 5월9일부터 주말(A)마다 촛불(D)집회(C)에 **참석(B)**해 온 서울 ㅅ고 신정아(18·사진)양은 "주말(A)마다 밤늦게(B)까지 촛불(D)집회(C)에 **참가(B)**하느라 몸(A)이 힘들(A)기도 하지만 마음(A)은 즐겁(A)기만 하다"며 "야간(B)자율(C)학습(B) 때문에 평일(A)에 집회(C) **참여(B)**를 못하(A)는 것이 아쉽다(B)"고 말(A)했다(A).
우연히 촛불(D)집회(C) 현장(B)에서 시민(A) 자유(A)발언(D)을 듣(A)고 집회(C)에 적극적(B)으로 **참여(B)**하게 됐다는 신양(B)은 마음(A)이 맞(A)는 청소년(A)들과 '10대(C) 연대(D)'라는 온라인(B) 모임(A)을 만들(A)어 지난 5월17일과 6

# 코퍼스 연구의 예 [2/3]

- 홍정하 · 김문조(2018), 신문 코퍼스에 나타나는 어휘 빈도의 시간적 변화는 어떻게 사회적 관심사를 반영하는가?



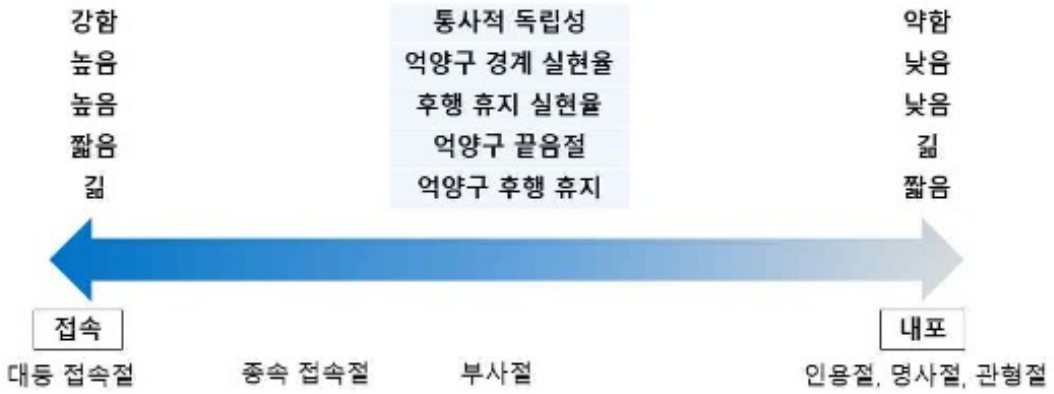

[그림 1] 인기 신기술 상품 품목 어휘 표현 및 연관어 분포 추세




[그림 9] '마음안식, 혼자 놀기, 사회인 야구, 아이돌' 관련 표현의 빈도 추세

# 코퍼스 연구의 예 [3/3]

- 전태희·신지영(2020), 절의 통사 유형에 따른 운율적 실현 양상
  - 구어의 운율과 문법 사이의 상관관계를 살핀 연구
  - 독백 자유 발화 음성 코퍼스 활용
  - 억양구 경계 실현(끊어 읽기)은 무작위적으로 이루어지는 것이 아니라 발화의 통사 구조와 밀접한 연관이 있음




<그림 4> 절 유형에 따른 통사 및 운율적 특성