# 인공지능과 시 읽기:
# 기계 해석과 인간 이해의 접점

김민서 (연세대학교)

## Word Embedding

컴퓨터가 인간의 언어를 이해하고 사용하는 방식은 워드임베딩 기술에 근거한다. 워드임베딩이란 단어를 숫자 배열(vector)로 변환하는 방법으로, 컴퓨터가 인간의 언어를 해석하고 처리할 수 있도록 돕는 자연어 처리 기법 중 하나다. 이 기술은 한 단어의 의미를 개별적으로 규정하기보다는, 해당 단어가 다른 단어들과 맺는 관계를 기반으로 정의한다는 관점에 기반한다. 워드임베딩은 보통 단어를 고차원 공간에서 밀집된(low-dimensional, dense) 벡터로 표현하며, 단어들 간의 관계(의미적 유사성 또는 문맥적 관련성)를 벡터 공간에서 상대적인 위치로 나타낸다. 이를 통해 단어뿐 아니라 문장, 문단, 또는 더 큰 규모의 텍스트도 벡터로 표현할 수 있다.

## Cosine Similarity

두 표현이 얼마나 가까운지, 즉 의미상 얼마나 유사한지 확인하기 위해 코사인 유사도(cosine similarity)를 계산한다. 코사인 유사도(cosine similarity)는 두 벡터 간의 방향적 유사도를 측정하는 방법으로, 두 벡터가 이루는 각도의 코사인 값을 이용한다. 수학적으로 코사인 유사도는 두 벡터의 내적(inner product)을 각 벡터의 크기(norm)로 나눈 값으로 계산된다. 이 값은 -1에서 1 사이의 범위로 나타나며, 값이 1에 가까울수록 두 벡터가 유사한 방향을 가짐을 의미한다. 0에 가까울수록 무관하고, -1에 가까울수록 반대 방향을 갖는다. 텍스트 데이터를 벡터 공간 모델로 변환한 후, 각 벡터 간의 코사인 유사도를 계산함으로써 문서 간의 유사성을 정량적으로 평가할 수 있다.

**시적 의미를 계량화하거나 수치화할 수 있는가?**
**계량된 결과는 인간의 이해와 해석의 범주 안에서 의미를 얻을 수 있는가?**

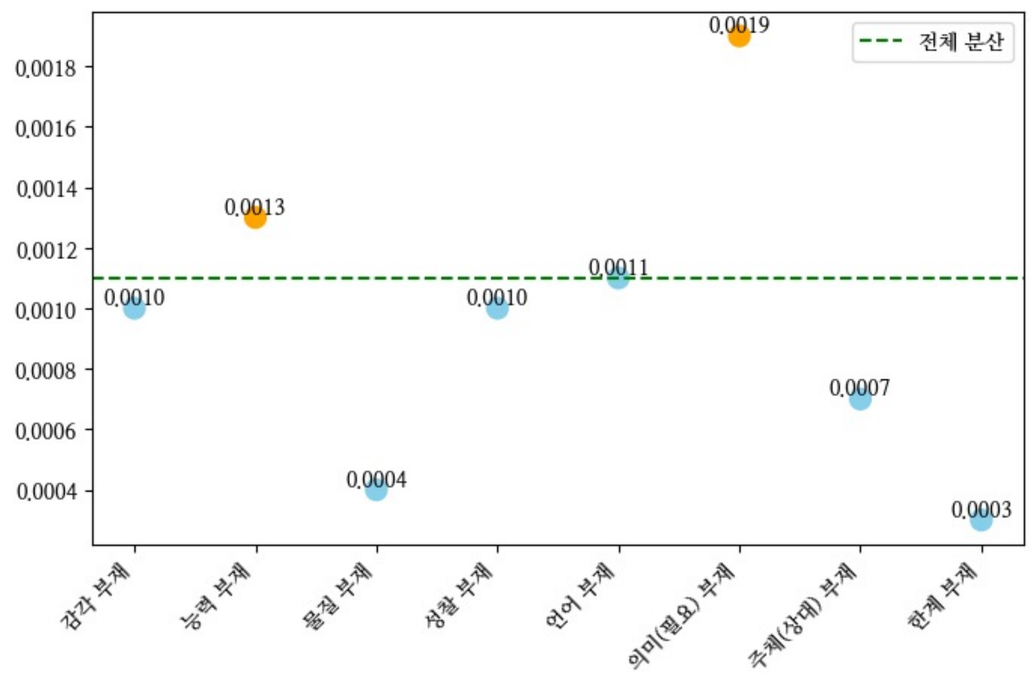
김수영 시에서 부재(absence)의 정도를 기계적으로 측정 → 부재의 정도가 높게, 혹은 낮게 측정되는 작품 확인 → 계량화의 의미를 구체적으로 해석

## 연구 과정 및 결과

1. 형용사 '없다'의 어근 '없'이 등장하는 시 100편 선정
2. 100편의 시에서 '없'다고 표현된 부재의 대상을 추출하고 유형화

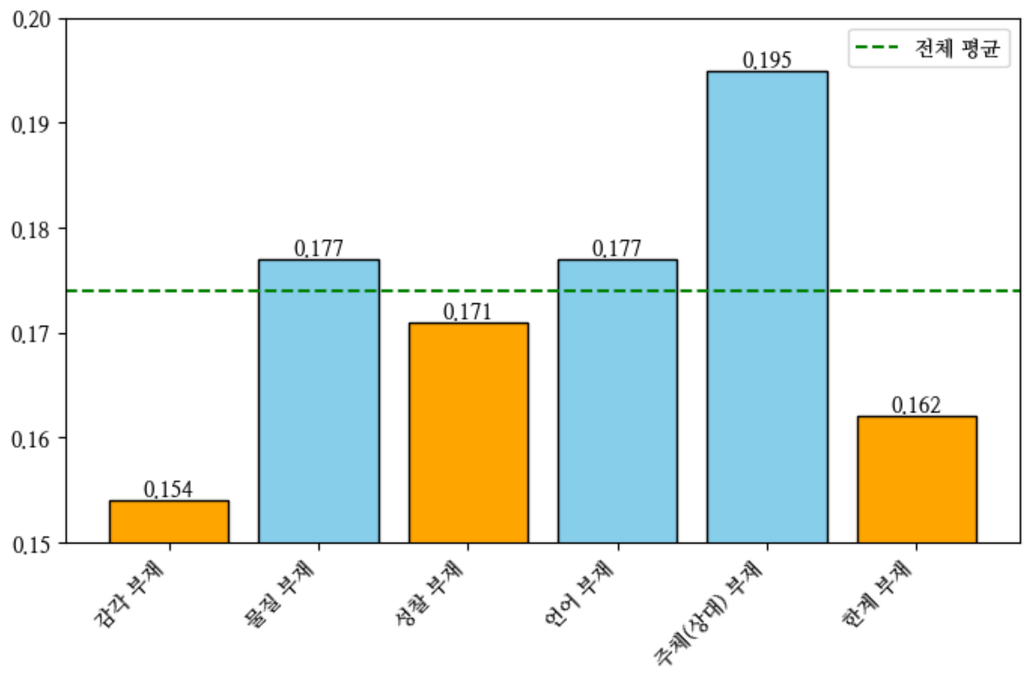
| | 부재의 대상에 따른 유형 구분 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 감각 | 능력(기력) | 물질 | 주체(상대) | 성찰 | 언어 | 의미(필요) | 한계 |
| 부재의 대상 및 구문 | 소리 | ~할 수 (없다) | 잔 | 너, 당신 | 찾다 | 어처구니(없다), 어이(없다) | 의미 | 한(없이) |
| | 빛 | ~한 일 (없다) | 돈 | 나 | 미련, 회한, 동요, 설움 | 높이, 폭, 규정, 표준 | 이유, 목적, 필요, 소용 | 수(없이) |
| | | 힘, 생기, 맥 | 지휘편 | ~한 사람 | 헤아리다 | 터무니(없다) | 쓸데(없다), 보잘것(없다), 덧(없다) | |
| | | | 집, 은거할 곳, 바닥 | 정체 | 정신 집중, 반성, 숙련 | 말, 이름, 질문, 약속 | 성과, 효과, 값 | |
| | | | 냉방장치 | 적 | 중용, 중립 | ~할 리 (없다) | | |
| | | | 타구 | 그림자 | | | | |
| | | | | 에미 | | | | |
| 작품 수 | 10 | 23 | 12 | 13 | 12 | 10 | 16 | 4 |

어학적 분류 기준

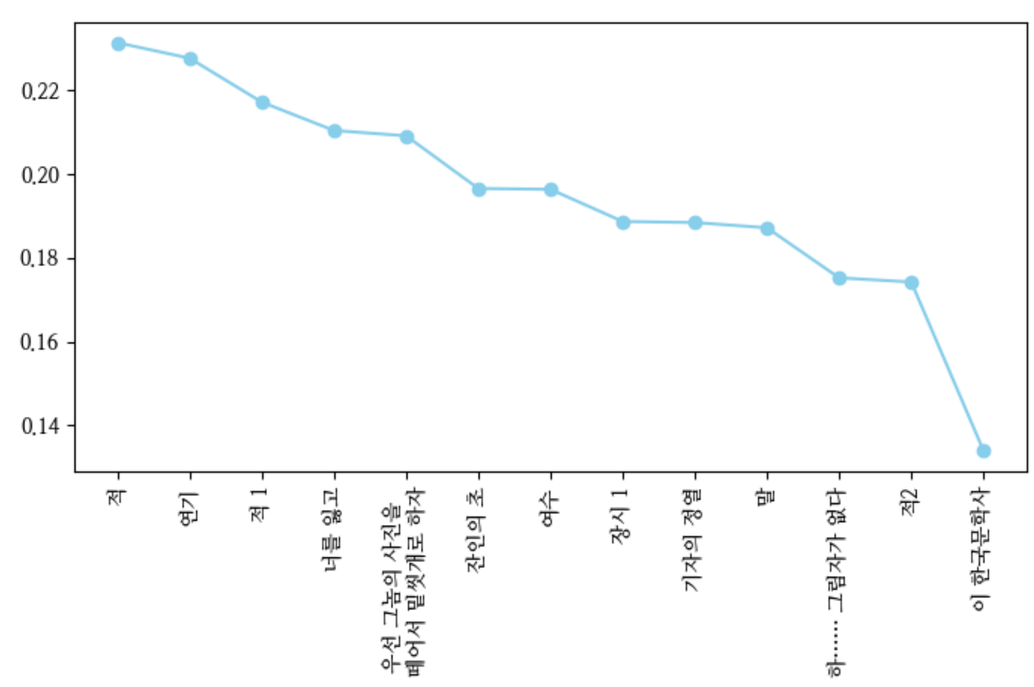



코사인 유사도 유형별 분산

3. 100편의 시를 임베딩하여 벡터화하고 어근 '없'과의 코사인 유사도 측정
4. 유형별 코사인 유사도의 분산을 분석하여 어학적 분류 기준과의 일치 여부 검토
5. 가장 높은 평균 코사인 유사도를 가지는 유형 도출: 주체(상대) 부재




코사인 유사도 유형별 평균




주체(상대) 부재 유형의 작품별 코사인 유사도

「적」 전문과 부재의 정도 해석

「이 한국문학사」 전문과 부재의 정도 해석