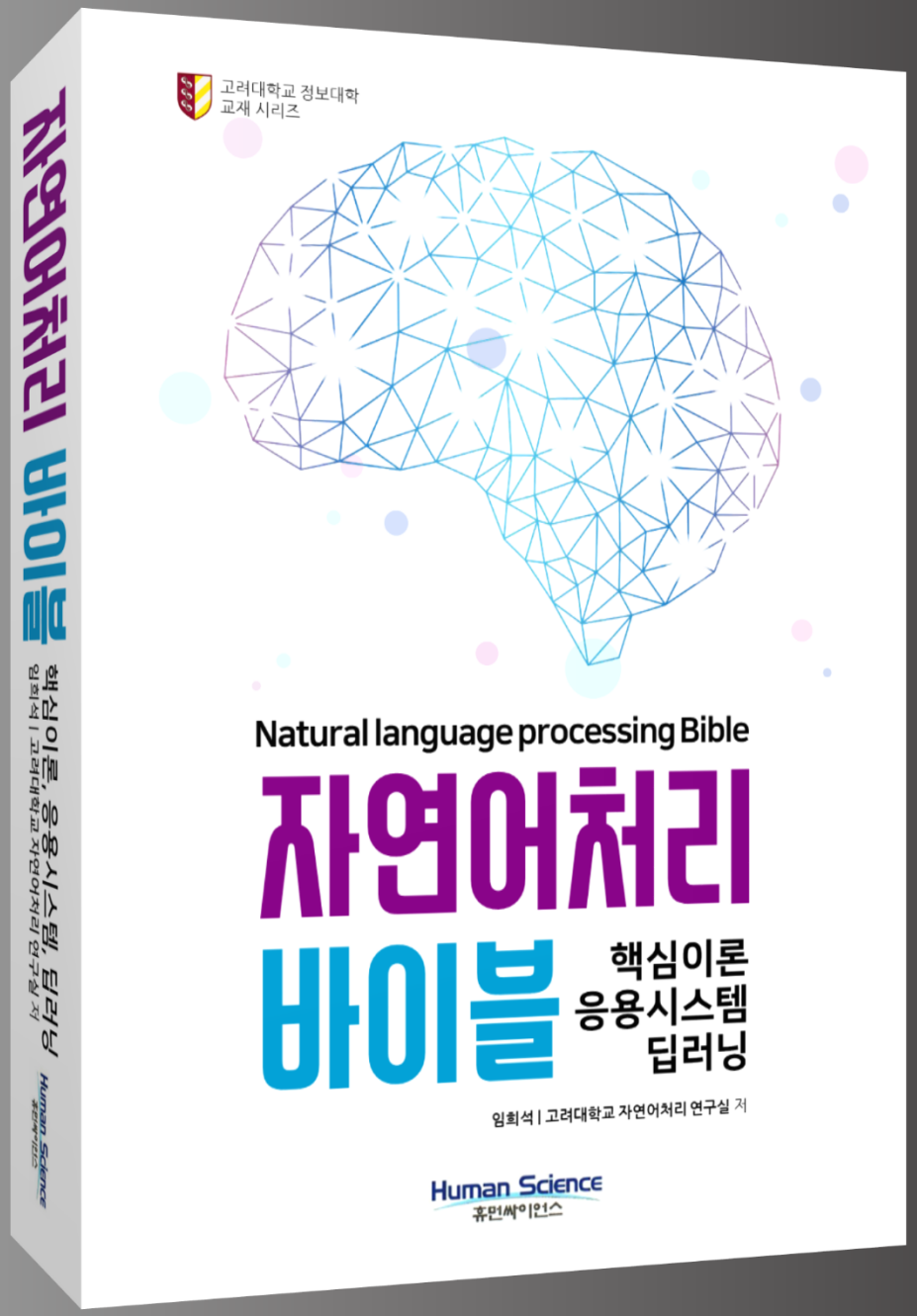


CHAPTER 15
문서 요약
(Text Summarization)

문서 요약이란?

- 문서요약이란 텍스트의 의미를 유지하면서 텍스트의 내용을 간략하게 줄이는 것임

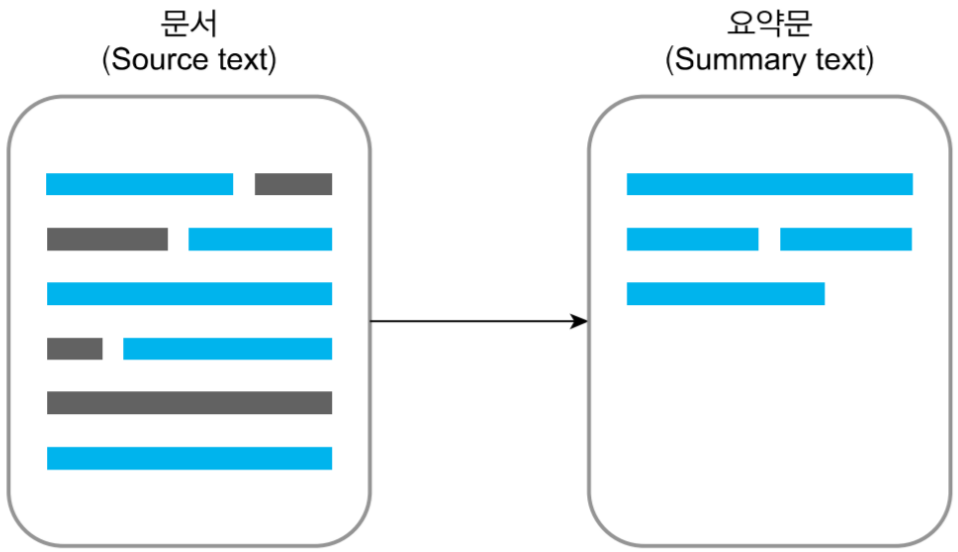


그림 15-1 Text summarization의 정의

문서 요약이란?

- 문서 요약은 I(interpretation), T(transformation), G(generation)으로 세 가지의 기본 과정으로 정의함
  - I : 주어진 문서 텍스트를 해석하여 컴퓨터가 이해할 수 있도록 표현하는 것
  - T : 문서 표현(source representation)을 요약문으로 표현될 수 있도록 변형/가공하는 것
  - G : 요약문에 대한 표현(summary representation)으로부터 최종 요약문을 생성하는 것

문서 요약 방법

- 문서 요약의 유형은 입력 문서, 출력 문서, 목적에 따라 다음과 같이 분류할 수 있음

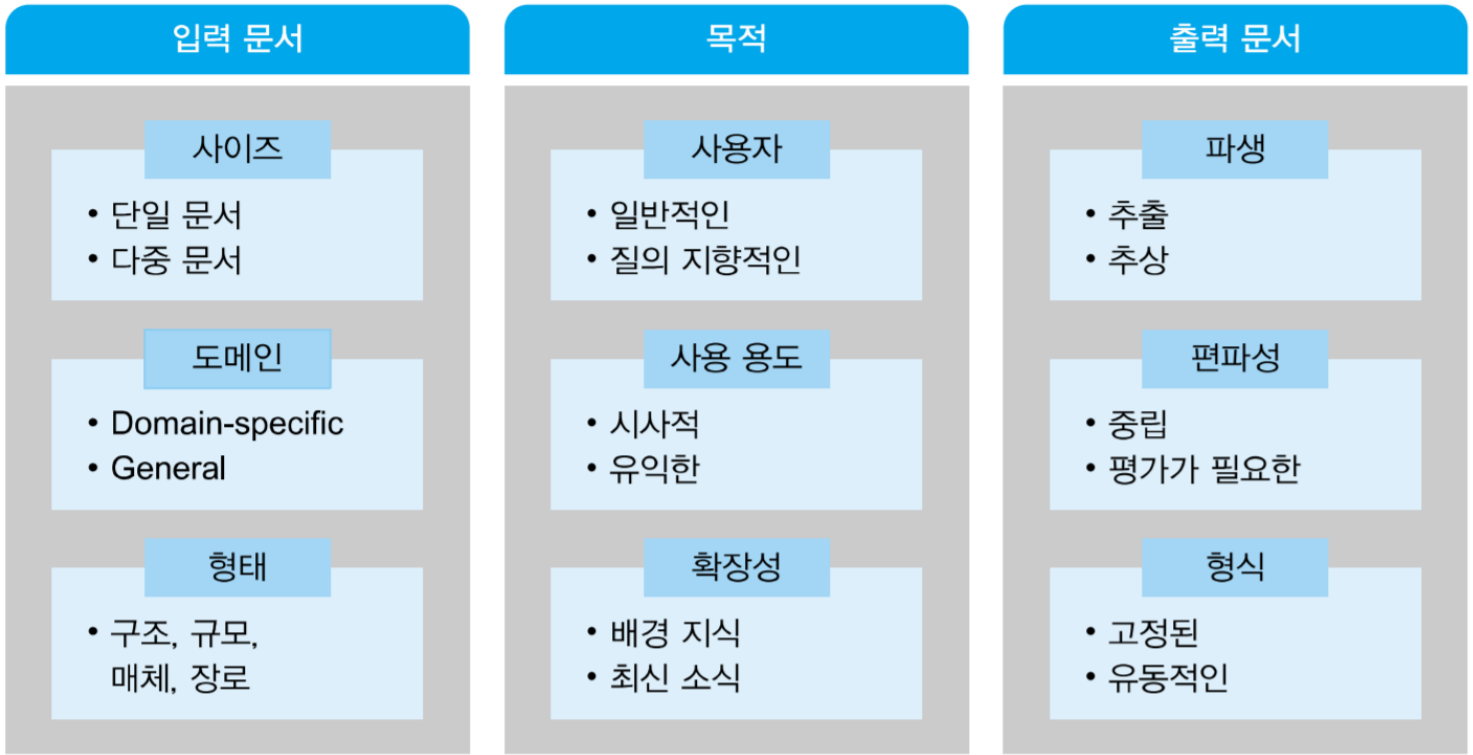


그림 15-2 여러 가지 기준에 따른 요약 시스템의 분류

문서 요약 방법

- 입력 문서
  - 입력 문서에 따라 요약 시스템이 다룰 수 있는 기준은 크게 세 가지로 분류할 수 있음
    - 문서의 크기 : 요약 시스템은 단일 입력 문서 또는 다중 입력 문서를 입력 문서로 가질 수 있음
    - 문서의 도메인(domain) : 문서 요약은 특정 도메인에 대한 요약을 생성하거나 일반적인 도메인의 요약을 생성할 수 있음
      - 예시) 특정 도메인 : 바이오메디컬(biomedical) 요약 시스템, 법률 문서 요약 시스템
    - 문서의 형태(form) : 입력 문서는 구조(structure), 규모(scale), 매체(mediums), 장르(genre)에 기반하여 여러가지 형태를 가질 수 있음
      - 예시) 구조 : 두괄식, 중괄식, 미괄식 / 규모 : 문서의 길이(트윗, 뉴스기사) / 매체 : 텍스트, 이미지, 비디오, 오디오 요약 / 장르 : 뉴스, 인터뷰, 리뷰, 소설

문서 요약 방법

- 목적
  - 사용자 : 단순히 입력 문서에 대한 요약이 아닌, 사용자가 알고 싶어하는 정보의 요약이 필요함
    - 질의 지향(query-oriented) 요약 시스템 : 사용자의 선호도를 고려한 요약 시스템
    - 일반적인(generic) 요약 시스템 : 입력 문서에 기반한 중요 정보를 보존하는 시스템

  - 사용 용도
    - 유용한 정보(informative)가 반영된 요약문 : 원본 문서의 필수적인 정보를 담고 있음
      - 예) 논문에서의 요약(abstract)
    - 시사적(indicative) 요약문 : 유익한 정보를 포함하기보다는 원본 문서의 전반적인 설명만을 포함함
      - 예) 책에서 머릿글 페이지의 요약문

  - 확장성 : 생성된 요약문은 원본 문서의 배경(background)에 중점을 두거나 일부 과거 문서들과 비교하여 최신 소식을 제공할 수 있음
    - 예) 코로나 바이러스와 관련된 최신 소식

문서 요약 방법

- 요약 내용의 형태에 따른 분류
  - 추출 요약 (Extractive)
  - 추상 요약 (Abstractive)

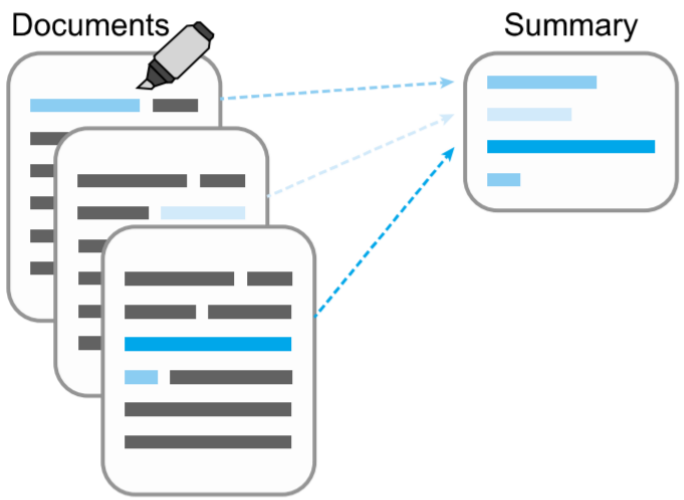


그림 15-4 추출 요약 예시

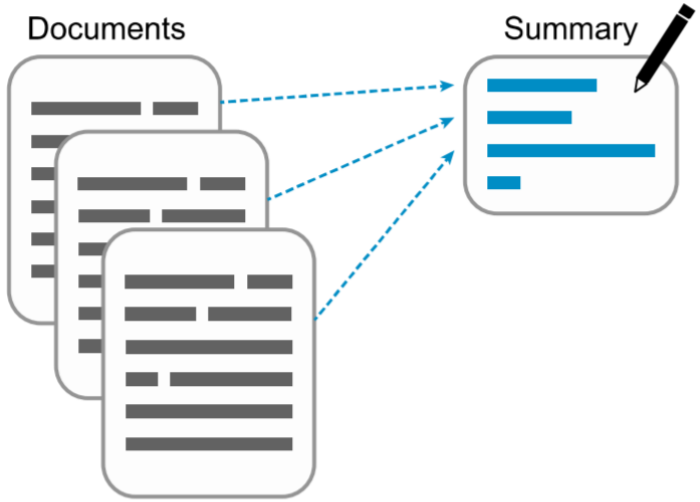


그림 15-5 추상 요약 예시

접근법

- 자동 요약 시스템을 개발하기 위한 다양한 접근법
  - 정보검색과 통계적인 기술을 이용한 방법
    - 토픽 표현 접근법
  - 기계학습 기반의 방법

통계적 접근법

- 문서 요약에서 통계적 접근은 문서의 주요 주제나 사용자의 요청 사항에 대해 일부 자질들을 이용하여 텍스트간(일반적으로 문장단위)의 연관성 점수를 계산하는 방식으로 요약문을 구성함
- 여기서 나오는 연관성 점수들 중 최고 점수를 받은 문장들이 요약문으로 사용됨
- 통계적 접근법에서 사용되는 자질들(features)
  - 용어 빈도(term frequency)
  - 텍스트의 위치
  - 잠재 의미 분석(latent semantic analysis, LSA)

통계적 접근법 – 주요 자질

- 용어 빈도 : 문서내에 빈번하게 사용되는 용어들의 중요도를 계산
  - $idf(t)$: 용어 t에 대한 idf 수식, $|D|$: 코퍼스 D에서 문서들의 개수, $|d/tNd|$: t를 포함하고 있는 문서들의 개수

$$idf(t) = \log \frac{|D|}{|d/tINd| + 1}$$

통계적 접근법 – 주요 자질

- 위치 정보
  - 텍스트의 위치 : 텍스트에서 단어들과 문장들의 위치는 문서내에 중요한 정보를 파악하기에 잠재적인 조건을 가지고 있음
  - 문서의 앞 또는 뒤쪽에 출현한 문장은 중요한 문장일 가능성이 있음
  - 각 문장에 대해 위치에 기반한 점수로 환산함
    - DP(direct proportion) : 처음 출현 또는 마지막 출현에 따른 점수 함수
    - IP(inverse proportion) : 위치 인덱스를 활용한 함수
    - GS(geometric sequence) : 모든 단어를 반영한 해당 단어의 점수를 도출하는 함수
    - BF(binary function) : 단어의 첫 등장에 대한 가중치 함수

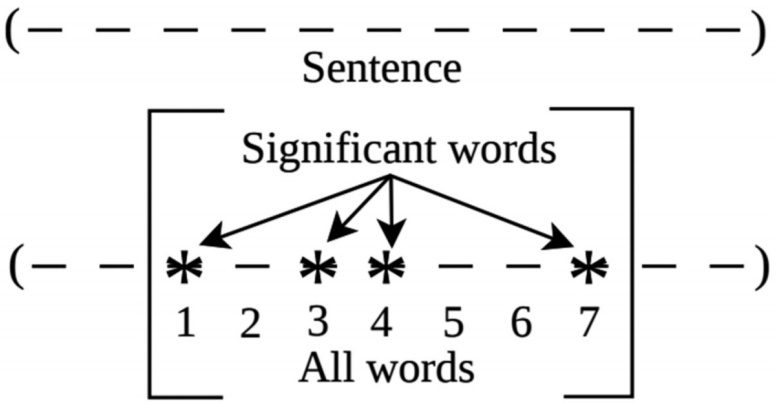


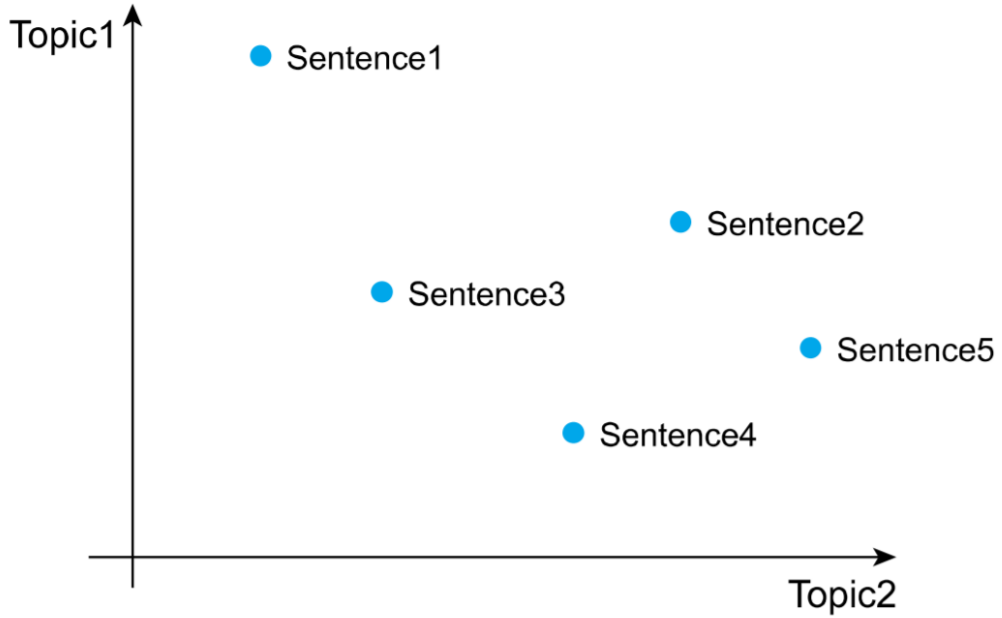
그림 15-6 단어 위치를 사용한 Luhn score[14]

$$Score(s) = \sum_{w_i INs} \frac{\log(freq(w_i)) * pos(w_i)}{|s|}$$

$|s|$: 문장의 길이
$pos(w_i)$: $DP\ or\ GS\ or\ BF$
$freq(w_i)$: 단어 w_i의 빈도

통계적 접근법 – 주요 자질

- 잠재 의미 분석 : 문서들과 문서들에 포함된 용어들 사이의 관계성을 탐색하는 분석 방법



	Topic1	Topic2
Sent1	0.9	0.25
Sent2	0.7	0.7
Sent3	0.45	0.4
Sent4	0.3	0.6
Sent5	0.4	0.9

그림 15-7 LSA기반의 문장 선택 예시[17]

기계학습 접근법

- 일반적으로, 기계학습 접근법은 다른 접근법들과 같이 사용하여 성능을 향상시키는 방향으로 활용됨
- 앞선 통계적 접근법의 자질들과 결합하여 문제를 해결하는 방향으로 결정 함수의 학습을 진행함
- 결정 함수 : 문서와 문서에 대한 추출 요약이 주어진 코퍼스에서의 기계학습 알고리즘 활용은 주로 문장이 요약문에 포함될지 아닐지를 결정하는 문제를 해결함
  - 베이지안 분류(bayes classification)

s_i: 문장 , $\vec{f}$: 자질 벡터
S: 요약문
$P(s_i INS)$: 상수값
$P(f_j|s_i INS), P(f_j)$: 코퍼스로부터 추정된 값
$P(s_i INS|\vec{f})$: s_i가 요약문에 들어갈 확률

$$P(s_i INS|\vec{f}) = \frac{\prod_{j=1}^{|\vec{f}|} P(f_j|s_i INS) * P(s_i INS)}{\prod_{j=1}^{|\vec{f}|} P(f_j)}$$

평가

- 요약 시스템에서의 평가는 정확해야 하며, 평가하는데 시간이 너무 많이 걸려서도 안됨
- 따라서, 사람이 평가하는 방법보다는 자동적으로 수행하는 알고리즘을 이용한 평가를 많이 사용함
  - ROUGE(Recall-Oriented Understudy for Gisting Evaluation)는 정답 요약과 모델이 생성한 요약을 비교해서 자동적으로 요약 시스템의 성능을 측정함
  - ROUGE의 알고리즘은 시스템이 만들어낸 요약과 정답 요약, 그리고 recall 값에 대한 n-gram의 개수를 계산하는 방식임

$$ROUGE-(N) = \frac{\sum_{S \in \sum m_{ref}} \sum_{N-gramINS} Count_{match}(N-gram)}{\sum_{S \in \sum m_{ref}} \sum_{N-gramINS} Count(N-gram)}$$

N: $N-gram$의 사이즈, $count(N-gram)$: 정답 요약에 있는 N-gram의 개수
$count_{match}(N-gram)$: 후보요약과 정답 요약에서 찾은 N-gram의 개수