텍스트 마이닝을 활용한 언어 자료의 분석과 연구

서울대 보건대학원 조원광

순서

언어 자료 활용 연구와 관련 통계 모델

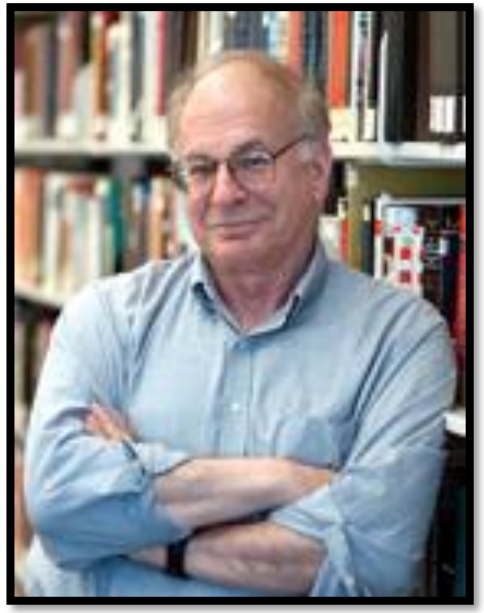
자연어 활용 연구의 동기와 장점

- 많은 분야에서 언어 자료 다루어 왔음
- 관심 현상이 인간 언어로 표현되는 경우 다수
- 집단 심리, 지식, 문화, 담론, 망탈리떼, 가치 체계 등
  - 인간의 행동과 판단에 영향을 미치는 구조적 변수들
  - 건강 등 신체 상태에도 영향을 미치는 경우 적지 않음 (집단 정서 -〉 스트레스 -〉 건강 변화)
- 막스 베버의 독일 언론 분석 발표 (Dickinson, 2013) (Evans & Aceves, 2016)

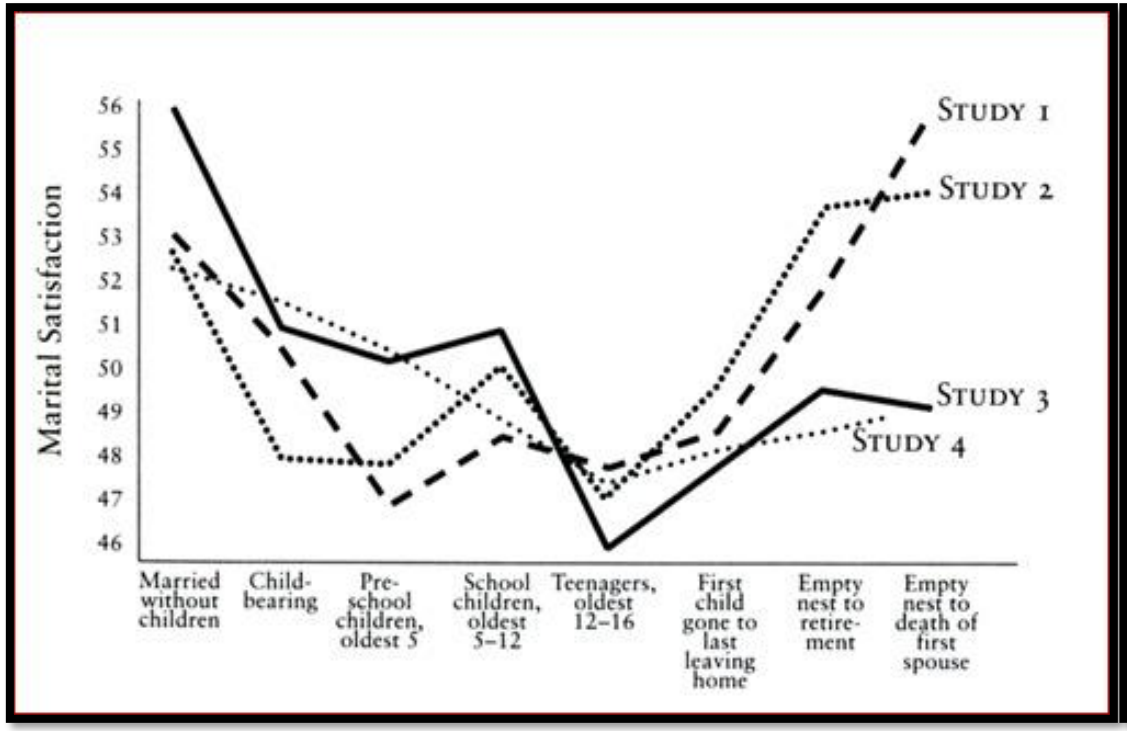
Public Domain, https://commons.wikimedia.org/w/index.php?curid=156949

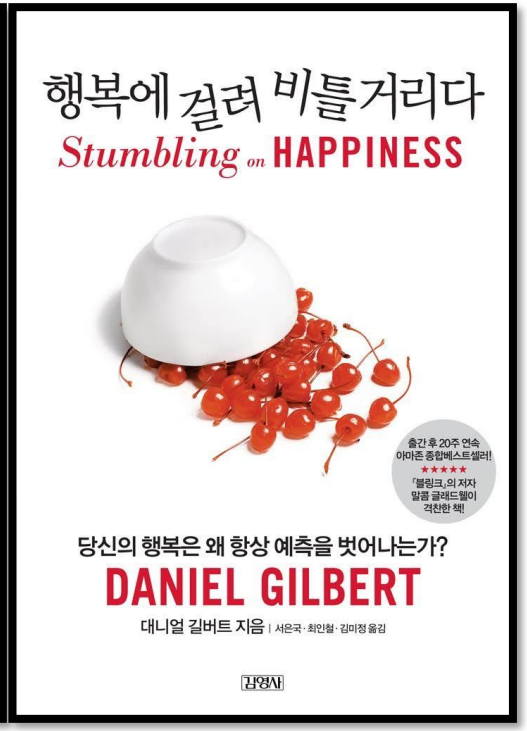
자연어 활용 연구의 동기와 장점

- 자연어(Natural Language): 인간이 일상적으로 쓰는 언어 (인공언어와 대비 / 엄밀한 정의는 아님)
- 자연어 자료의 장점 (1): 조사자의 프레임과 응답자의 회고 효과에서 자유로운 자료 다수
  - 자발적으로, 실시간으로 작성되는 자료인 경우 다수



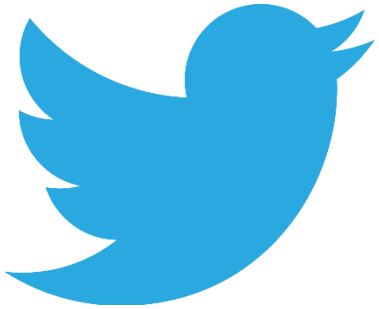
By https://nihrecord.nih.gov/newsletters/04_13_2004/story02.htm, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1266697





자연어 활용 연구의 동기와 장점

- 자연어 자료의 장점 (2): 디지털 시대에 종류와 양이 급증
  - 기록의 시대
  - 수집 방법 다양: 인터넷 스크래핑, 설문 조사에서 개방형 질문 등



- 자발적으로 작성된 언어 자료로 접근하면 유리한 변수들 다수
  - 집합적 심성 / 정서 / 태도 / 프레임 / 지식

- 해당 변수 현황 파악 + 타 변수의 설명 변수로 활용되는 경우 다수
  - e.g., 주가 예측, 특정 제품에 대한 태도 판단

자연어 활용 연구의 동기와 장점

- 자연어 자료 활용의 난점: 비정형성

평가자	A식당의 맛	A식당의 서비스	A식당의 청결
1	4	3	3
2	3	4	4
3	5	5	3
4	6	7	6
5	2	2	1

평가자 1	그 집 족발 상당해~ 장난 아니야~ 사장님 손가락에 금붙이 장난 아닌거 봤지? 옆 식당 다 먹어서 벽트고 확장했다니까? 홀만 300평이잖아!
평가자 2	A 식당의 족발은 마치 5월의 푸른 햇살 같습니다. 한 입 베어 물면, 아직 완전히 익지 않은 것이라 착각할 만한 쫄깃쫄깃한 살코기에 정신이 번쩍 들면서 마음이 맑아지거든요.
평가자 3	거기 뭐 그냥 그렇지. 나쁘다는 건 아니지만, 그렇다고 뭐 줄 서서 먹을 정도는 아니고....내가 전에 가봤는데, 사장님이 인자하고 친절하시더라고. 요리하는 과정도 보이고. 그런데 뭐 또 갈 수도 있고 아님 뭐 다른데 가도 괜찮고.

- 전자는 처리가 간단하지만, 후자는 분류부터 큰 도전
- 풍부한 정보를 가지고 있지만, 이를 추출하기 쉽지 않음: 언어 자료가 빅데이터라 불리는 이유
- 다양한 과정을 거쳐서 정보 추출

자연어 활용 연구의 동기와 장점

- 여러 분야에서 언어 자료로 연구하는 전통적인 방법
  - (1) Interpretivist text analysis (2) Systematic qualitative coding (Kozlowski, Taddy, & Evans, 2018)

By Exeter Centre for Advanced International Studies Research Priorities, Fair use, https://en.wikipedia.org/w/index.php?curid=23182200

- (1) Interpretivist text analysis
  - 연구자가 전체적으로 깊이 텍스트를 읽고 거기에 존재하는 의미와 구조 파악
  - 재생 가능성에 한계 -> 연구자 주관성 개입이라는 비판 피하기 어려움
  - 비판하지 않더라도, 따라하기가 어려움

- (2) Systematic qualitative coding
  - 미리 주제를 정하고, 해당 주제 혹은 요소가 텍스트에 존재하는지 판단
  - 여러 연구자들의 판단의 일관성 혹은 신뢰성 테스트 (e.g. kappa)
  - 연구 질문이 복잡할 때 쓰기 어려움. 방법적 한계 보고 (저빈도 범주 포함할 때 신뢰성 테스트 타당성 저해 등) (Viera & Garrett, 2005)
  - 탐색적 분석을 할 때 쓰기 어려움 (미리 초점이 있어야 함)

자연어 활용 연구의 동기와 장점

- 무엇보다, 둘 다 대량 자료 처리가 불가능

- 만약 연구나 작업의 목표 상, 다루어야 하는 자료가 특정 영역 '전체'로 잡혀 있거나 (예를 들어 특정 주제에 대해 디시인사이드에서 논의된 내용 전체) 자료 크기가 인간이 다루기 힘들 정도로 많다면, (개방형 질문 응답 5만건), 전통적 방법을 쓰기 힘들고 자동화된 방법을 활용해야 함

- 연구의 목표가 언어 자료에 존재하는 특이한 구조를 탐색적으로 잡아내는 것이라면, unsupervised learning methods로 분류되는 기법들이 유용 (e.g., semantic network, topic model, vector space embedding model)

자연어 대상 통계 모델의 분류

- (1) 등장 단어의 종류와 빈도 추출
  - Danner, D. D., Snowden, D. A., Friesen, W. V. (2001). Positive emotions in early life and longevity: Findings from the nun study. *Journal of Personality and Social Psychology*, 80, 804-813.

Table 1
Reliability of the Emotion Coding as Indicated by the Number of Emotion Words and Sentences Scored by Two Coders for Autobiographies Written in Early Life by 180 Participants in the Nun Study

Unit of analysis and emotion	Count		Correlation of counts		
				Final coding and each coder[a]	
	Coder A	Coder B	Coders A and B[a]	Coder A	Coder B
Words					
Positive	1,243	1,242	.96 (.95, .97)	.99 (.98, .99)	.98 (.98, .99)
Negative	206	192	.89 (.85, .94)	.97 (.94, .99)	.94 (.91, .97)
Neutral	16	17	.78 (.64, .93)	.97 (.90, 1.00)	.82 (.69, .95)
Sentences					
Positive	1,006	1,017	.97 (.96, .98)	.99 (.98, .99)	.99 (.98, .99)
Negative	196	179	.90 (.85, .95)	.97 (.96, .99)	.94 (.91, .97)
Neutral	16	17	.78 (.64, .93)	.97 (.90, 1.00)	.82 (.69, .95)

Note. For all correlations, $p < .0001$.
[a] 95% confidence intervals appear in parentheses.

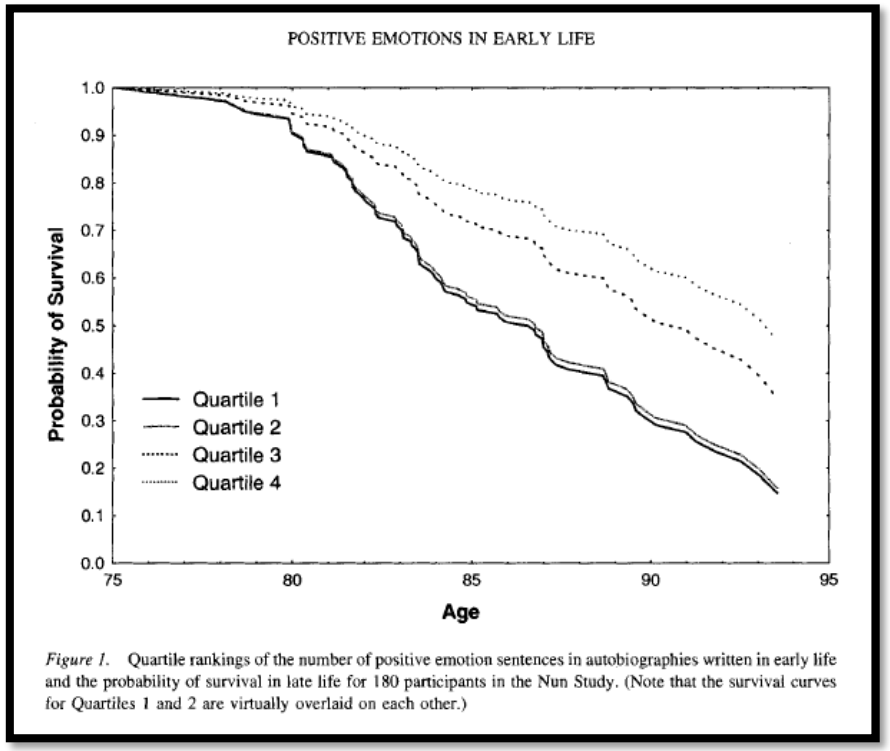
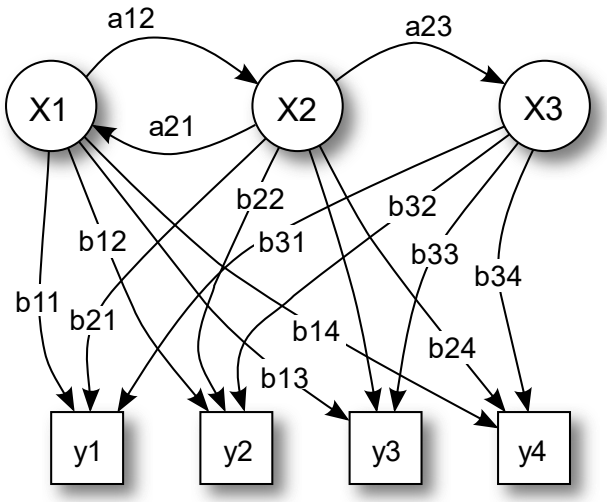


Figure 1. Quartile rankings of the number of positive emotion sentences in autobiographies written in early life and the probability of survival in late life for 180 participants in the Nun Study. (Note that the survival curves for Quartiles 1 and 2 are virtually overlaid on each other.)

자연어 대상 통계 모델의 분류

- (2) 등장 단어들 사이의 관계 (동시 출현, 순서, 위치 등)
  - 단어의 의미는 단일 단어가 아니라 단어들과의 관계에 의해 규정
    - “귀엽다”라는 단어의 의미는 해당 문장 전후에 어떤 단어가 위치하느냐에 따라 달라짐 (언어의 외부성)
  - 단어들의 관계는 의미만이 아니라, 개별 단어의 속성을 추정하는 중요 정보 -〉 첫 번째 정보의 강화
    - E.g., 문장 안에서 단어 출현 순서와 연결 정보를 활용하여 개별 단어에 품사 할당
    - E.g., 전후 단어와의 관계에 따라 (예를 들어 shifter의 존재 유무) 단어에 실린 정서 파악



자연어를 활용한 통계 모델 분류

Supervised learning

-지도 학습은 기본적으로 모델이 응답 변수를 설명하거나 예측하는 능력을 증가시키기 위해 모델을 학습시키는 방법을 의미

-자연어 자료에서 얻은 정보를 활용하여, 관심 응답 변수를 (주가, 정보의 진실성 여부) 설명하는 모델을 만드는 경우가 좋은 예

-다양한 알고리즘이 사용됨: 회귀분석, SVM, Tree method, deep learning, 등등

Unsupervised learning

-비지도 학습은 통계 모델을 '지도'할 응답변수가 없는 상태에서 변수 간, 혹은 케이스 간 관계를 이해하기 위한 통계 모델 / 즉 데이터 내에 존재하는 구조의 포착

-자연어 자료를 대상으로 하는 분석의 경우, 자료에 존재하는 여러 종류의 구조를 (주제, 가장 지배적인 단어간 연결 등) 포착하기 위해 사용

-자연어 자료의 경우, 다음 세 가지 종류의 비지도학습 통계 모델이 유명: (1) semantic network, (2) topic modeling, (3) vector space embedding

토픽 모델링을 활용한 연구 사례

토픽 모델링의 기본 논리

- 토픽 모델링
  - LSA -〉 PLSA -〉 LDA -〉 …
  - LDA: 다수의 문서에서, 해당 문서를 가장 잘 설명하는 복수의 단어 확률 분포와 (Topic) 문서 별 토픽 분포를 추출하는 기법
    - 개별 문서를 단순한 단어의 집합으로 (a bag of words) 판단: 단어의 종류 / 빈도 / 동시 출현 정보를 활용
    - 토픽은 전체 단어의 확률 분포라고 가정 (e.g., dog - 0.001, cat - 0.0015, cute - 0.002, food - 0.0001……)
    - 토픽은 복수로 존재할 수 있고, 개별 문서에 복수의 토픽의 특정 분포를 가지고 존재 (soft membership)
  - 주어진 문서들이 토픽들과 문서별 토픽 분포로부터 무작위 생성 (random generating) 되었다고 가정
  - 어떤 단어 확률 분포와 문서 별 토픽 분포가 가정되어야, 현 데이터를 생성할 개연성이 가장 높은가? -〉 통계적 문제로 질문을 전환 -〉 다량의 문서에 존재하는 주제들을 통계 모델로 접근할 가능성 확보
  - Latent Dirichlet Allocation (LDA)가 가장 유명하고 폭넓게 쓰이는 방법
    - 추정해야 할 파라미터들의 사전확률분포가 Dirichlet 분포라고 가정하는데서 온 명칭
  - 이후 CTM, DTM, STM 등으로 다양하게 발달

Topics

Documents

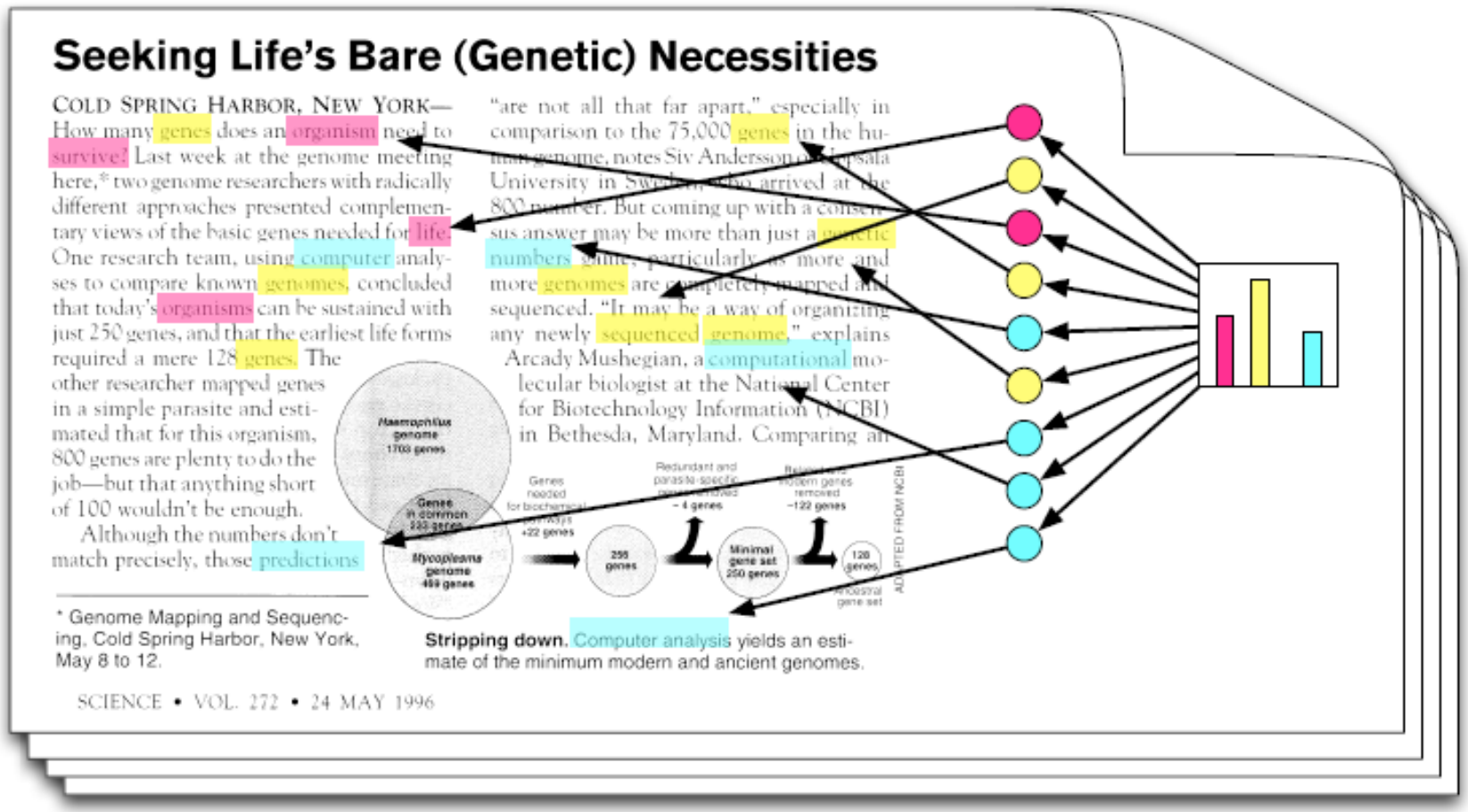
Topic proportions and assignments

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	



Seeking Life's Bare (Genetic) Necessities

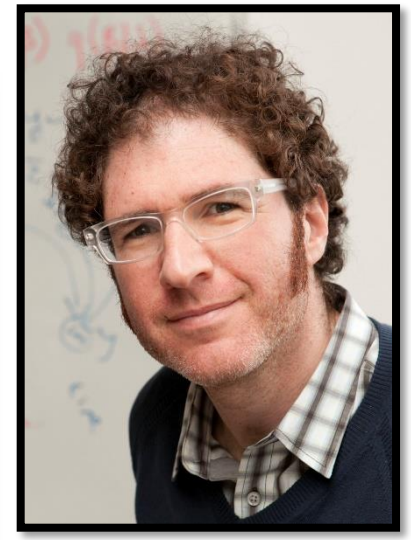
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

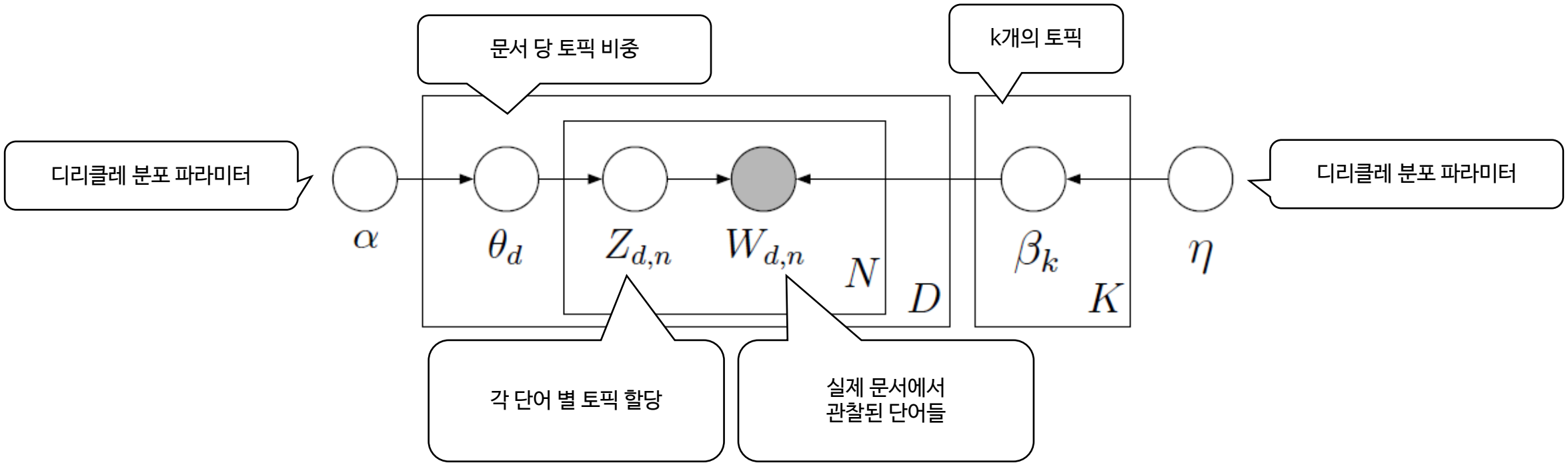
SCIENCE • VOL. 272 • 24 MAY 1996



David M. Blei

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77.
doi:10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications, 10*(71), 34.



$$p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}, \alpha, \eta)$$

$$\frac{p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}{\int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{\vec{z}} p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}$$

특히 분모 때문에 해가 intractable하므로, 근사적 기법 활용

활용 사례: 토픽 모델링과 담론 연구

- Topic modeling은 '담론 연구' 분야에 보충적 방법으로 여겨질 가능성이 있음
  - 기존에는 "Interpretivist text analysis" 가 담론을 연구하는 주된 방법이었음

- 담론이란 무엇인가? : 유의미한 주체, 대상, 그리고 개념간 관계를 규정하는 잠재적 구조
  - 시공간에 따라 달라짐 (Foucault, 2013)
  - E.g., 근대 임상 의학 -〉 유의미한 대상: 병소, 세포, 세균, 생명, / 유의미한 주체: 국가 공인 보건 인력 / 합리적 개념 연결망: biomedicine

- 담론은 언어 그 자체가 아니라, 그 뒤에서 그것의 질서를 규정하는 잠재적 구조

활용 사례: 토픽 모델링과 담론 연구

- 토픽 모델링의 결과는 (토픽들 + 개별 문서의 토픽 분포) 담론의 흔적을 포착하는 좋은 방법

- (1) 하나의 담론이 작동할 때, 문서에 자주 출현하는 단어와 단어의 조합을 남김
  - 토픽 모델링은 해당 특징을 단어 확률 분포로 포착 -〉 토픽으로부터 담론 역추론 가능

- (2) 담론은 단수가 아니라 복수로 존재 / 경쟁 혹은 연합하며 작동 + 동일한 담론도 다양하게 표현
  - 암에 대한 문서에는 임상 의학 담론만이 작동하는 것이 아니라, 국가의 건강 관리 담론 등도 함께 존재
  - 동일한 임상 의학 담론도, 감기를 다룰 때와 암을 다룰 때 다른 방식으로 표현
  - 토픽 모델링은 단순히 최빈 단어를 세는 것이 아니라 확률 분포를 추출하려고 시도하고, 이런 토픽을 복수로 가정함으로써, 전체에서 가장 강한 경향 만을 추출하는 대신 담론 작동의 다양한 복수성을 포착할 수 있음

활용 사례: 토픽 모델링과 담론 연구

- 담론은 잠재적이며 유동적인 구조 변수, 토픽 역시 관찰된 단어가 아닌 잠재 변수
  - 둘 다 명시적으로 표현되지 않지만, 잠재적 수준에서 명시적으로 표현되는 현상에 영향을 미침
  - 어떤 대상, 주체, 개념 간의 연결이 언어 자료에 등장하는지에 영향을 미침 -〉 해당 언어 자료에 유의미하게 다루어지는 것을 규정한다고 해석 가능

- 토픽모델링이 담론이라는 개념의 전체를 모두 포착한다고 볼 수는 없지만, 적어도 그것에 접근하는 하나의 도구이자 방법이 될 수 있으며, 연구자가 해석의 보완물 내지는 도구로 삼을 수 있음
  - Nelson의 3단계론: 1단계 컴퓨터로 패턴 추출 -〉 2단계 연구자의 해석과 서사 확립 -〉 3단계 해석과 서사를 확증하는 모델링 (Nelson, 2017)
- 담론 뿐만 아니라 [언어로 표현되는 집합적 인식의 잠재 구조]를 가리키는 변수 전반에 사용 가능
  - 집합적 정서, 프레이밍, 정보 체계 등

활용 사례: 토픽 모델링과 담론 연구

- 토픽 모델링의 종류 (다양한 응용과 발전)
- CTM (Correlated Topic Model)
  - 토픽 간의 상관관계가 있을 수 있음 (e.g., 민주주의 토픽은 탄핵 토픽과 함께 등장하는 경향이 있다)
  - 토픽의 문서별 비중을 추정할 때, 토픽 간의 상관관계가 존재할 수 있다고 가정하고 이를 함께 추정
- DTM (Dynamic Topic Model)
  - 토픽의 내부 구성이 시간에 따라 변화 가능 (e.g., 똑같은 민주주의 토픽이라도, 100년 동안 강조점이 달라질 수 있음)
  - 토픽의 내부 구성이 시간에 따라 달라진다는 점 가정하고, 이를 추정
- STM (Structural Topic Model)
  - 문서 내 토픽의 비중과 토픽의 구성이 문서의 각종 정보에 영향 받을 수 있음 (저자 카테고리, 작성 시점 등)
  - 다양한 문서의 메타 정보가 토픽 비중과 구성에 미치는 영향을 동시에 추정
- 담론의 변화, 담론에 영향을 미치는 요소들 등을 발굴하는데 활용 가능

토픽 모델링의 한계와 극복

- (1) 단어 간 관계 정보 직접 측정 및 추정 불가능
  - 동시 출현이 일종의 관계 정보이나, 그보다 세밀한 정보는 무시 (document = a bag of words)
- (2) 각종 하이퍼 파라미터, 특히 토픽 개수를 결정하는 확고한 방법이 없음 (Gerlach, Peixoto, & Altmann, 2018)
  - 현존하는 방법은 토픽 개수를 다양화하여, 각 토픽 모델의 질을 비교
  - 모델 질을 판단하는 지표에 대한 합의 미비 (held-out likelihood? Exclusivity?)
- (3) 사전 분포에 대한 이론적 정당화 미비 (Gerlach, Peixoto, & Altmann, 2018)
  - 예를 들어 디리클레 분포를 쓰는 이유는 그것이 다항분포의 conjugate prior라는 점이 상당부분
- (4) Short Text에서는 모델의 질이 떨어짐 (Quan, Kit, Ge, & Pan, 2015)
  - 인터넷에 존재하는 자연어 정보의 대부분이 short text / 이를 극복하려는 노력이 이어지고 있음

토픽 모델링의 한계와 극복

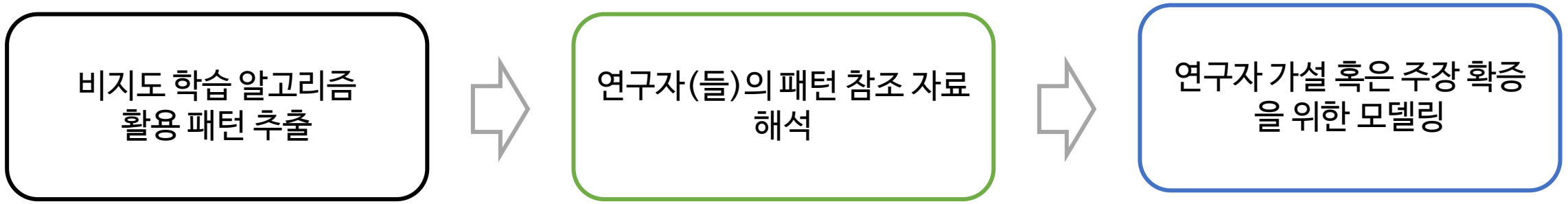
- 토픽 모델링 이외에도 언어 자료에서 해당 자료의 특징을 규정하는 일종의 '잠재적 변수'를 측정하려는 시도는 다양
- Example 1: Semantic network analysis
  - 언어 역시 네트워크 구조를 가지고 있음 (Newman, Barabasi, & Watts, 2011)
  - 다양한 네트워크 구조에서 파생하는 혹은 정의되는 정보를 추출하기 위한 시도 존재
    - 주요 단어/개념과 어떤 단어/개념이 연결되어 있는가?
    - 단어 간 관계를 고려했을 때, 어떤 단어/개념이 가장 중심적인가? 특정 역할을 하는 단어/개념이 존재하는가?
    - 다른 단어들에 비해 특히 자주 연결되어 응집적인 구조를 형성하는 단어 집합이 있는가? 이들은 무엇을 의미하는가?
    - 특정한 정서가 표출되는 대상과 그 대상의 속성은 무엇인가?
  - 연결 정보를 좀 더 다채롭게 활용할 수 있는 방법

토픽 모델링의 한계와 극복

- Example 2: Vector space embedding model
  - Vector space word embedding model 맥락 정보를 활용하여 단어를 벡터로 전환시키는 분석
    - 다양한 분석 가능 -〉 직접 연결되어 있지 않지만 구조적으로 유사한 단어를 추출하는 것이 좋은 예 (E.g., If “volleyball” and “softball” are located near the so-called “feminine words”, they could be categorized into so-called “feminine sports”, even though they do not co-appear in sentences (Kozlowski, Taddy, & Evans, 2018))
    - 단어를 가지고 여러 연산이 가능해지므로, 수많은 활용 사례가 존재할 수 있음

- 모든 방법은 한계와 단점을 가지지만, 적절히 연합하고 보충해서 사용하면 그것을 돌파할 수 있음

Nelson이 제안한 프로세스 (Nelson, 2017)



텍스트 마이닝 활용법에 대한 연구 필요성

1. 각종 텍스트 마이닝 결과물에 대한 이론적 논의 늘어날 필요
2. 더불어, 질적으로 차이나는 변수를 포착하기 위한 알고리즘 최적화 혹은 개발 고민 필요

- **사례: 의미망 네트워크에서 무엇이 가장 적절한 community detection 알고리즘인가?**

- 의미망 네트워크의 sub-community -> 해당 자료에 존재하는 세부 중요 의미망 포착에 매우 중요
  - 후보 알고리즘
  - (1) clique, clan등 전통적 알고리즘 -> 내외부의 상대적 응집성을 고려하지 않으므로 잘 사용되지 않음
  - (2) 세 가지 알고리즘 카테고리가 주로 동원: methods based on statistical inference, methods based on optimization, methods based on dynamics (Fortunato & Hric, 2016)

텍스트 마이닝 활용법에 대한 연구 필요성

- 기존 벤치마크의 난점 -〉 의미망 네트워크 혹은 의미 추출을 염두에 둔 실험의 부재
  - 통상 벤치마크 결과에 기반하여 알고리즘을 선택하나, 의미망 네트워크를 염두에 둔 결과는 드문편 (Yang, Algesheimer, & Tessone, 2016)
  - 하지만 의미망 네트워크의 커뮤티니는 나름의 독특함을 가짐 -〉 경로 거리 민감성 (“원숭이 엉덩이는 빨개” “빨간 건 사과” “사과는 맛있어” -〉 원숭이와 맛있어는 연결되어 있는가?)
  - 이 점을 염두에 두면, Algorithms based on dynamics (e.g., Walktrap, Infomap) 가 좋은 후보일 수 있음. 그것들이 벤치마크에서 가장 좋은 성과를 거두지 못하더라도
  - Resolution problem을 극복하면서도, 경로거리에서 오는 문제를 통제할 수 있기 때문 (Pons & Latapy, 2005)

References

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77. doi:10.1145/2133806.2133826
- Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications, 10*(71), 34.
- Danner, D. D., Snowdon, D. A., & Friesen, W. V. (2001). Positive emotions in early life and longevity: findings from the nun study. *Journal of Personality and Social Psychology, 80*(5), 804.
- Dickinson, R. (2013). Weber's sociology of the press and journalism: continuities in contemporary sociologies of journalists and the media. *Max Weber Studies*, 13(2). doi:10.15543/mws/2013/2/5
- Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*, 42(1), 21-50. doi:10.1146/annurev-soc-081715-074206
- Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics reports, 659*, 1-44. doi:10.1016/j.physrep.2016.09.002
- Foucault, M. (2013). *Archaeology of knowledge*: Routledge.
- Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science advances, 4*(7), eaaq1360.
- Kozlowski, A. C., Taddy, M., & Evans, J. A. (2018). The Geometry of Culture: Analyzing Meaning through Word Embeddings. *ArXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/#abs/2018arXiv180309288K
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists:* Springer Science & Business Media.
- Nelson, L. K. (2017). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*.
- Newman, M., Barabasi, A.-L., & Watts, D. J. (2011). *The structure and dynamics of networks* (Vol. 19): Princeton University Press.
- Pons, P., & Latapy, M. (2005). *Computing communities in large networks using random walks.* Paper presented at the International symposium on computer and information sciences.
- Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). *Short and Sparse Text Topic Modeling via Self-Aggregation.* Paper presented at the IJCAI.
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R package for structural topic models. *R package, 1*, 12.
- Viera, A. J., & Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine, 37*(5), 360-363.
- Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Sci Rep, 6*, 30750. doi:10.1038/srep30750