Natural language processing Bible

Basic

서수원

Business Intelligence Lab.

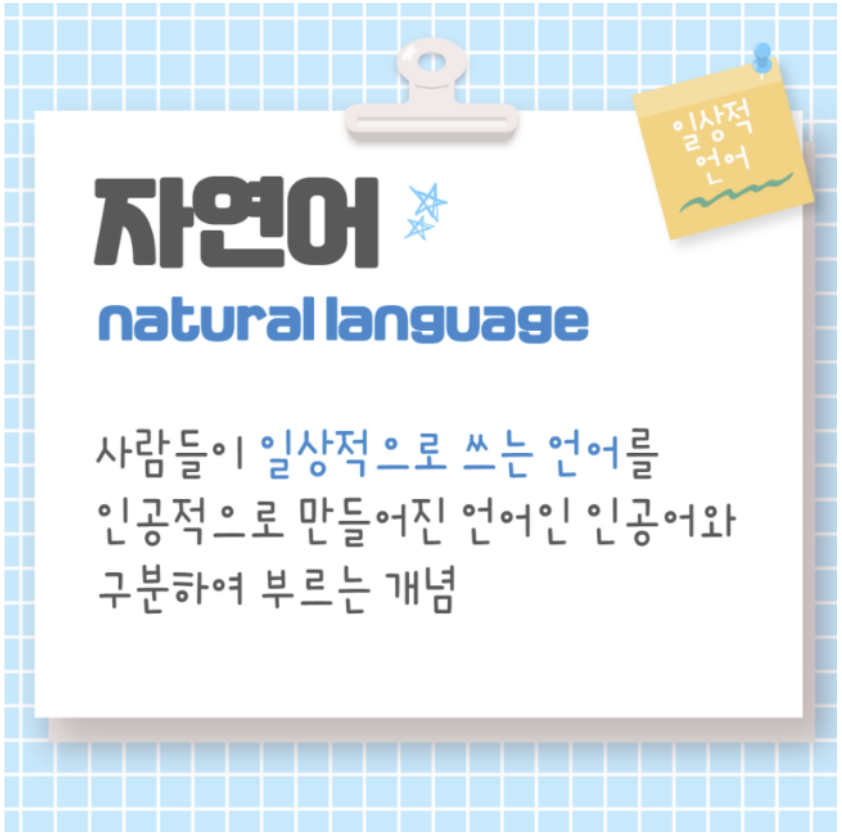
산업경영공학과, 명지대학교

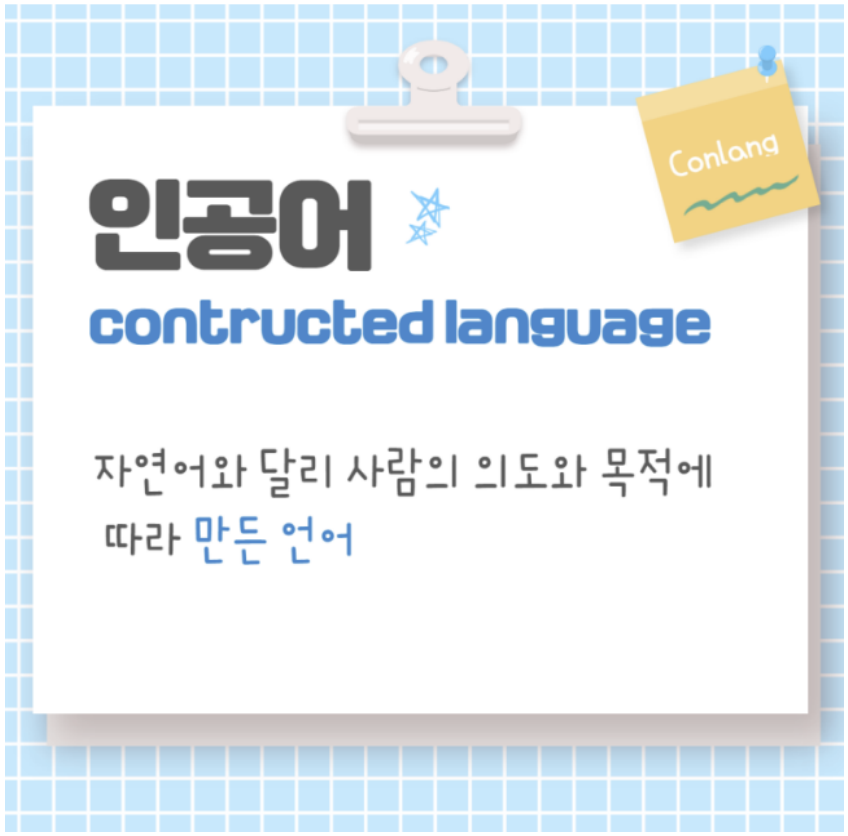
01

자연어처리의 기본

자연어와 인공어

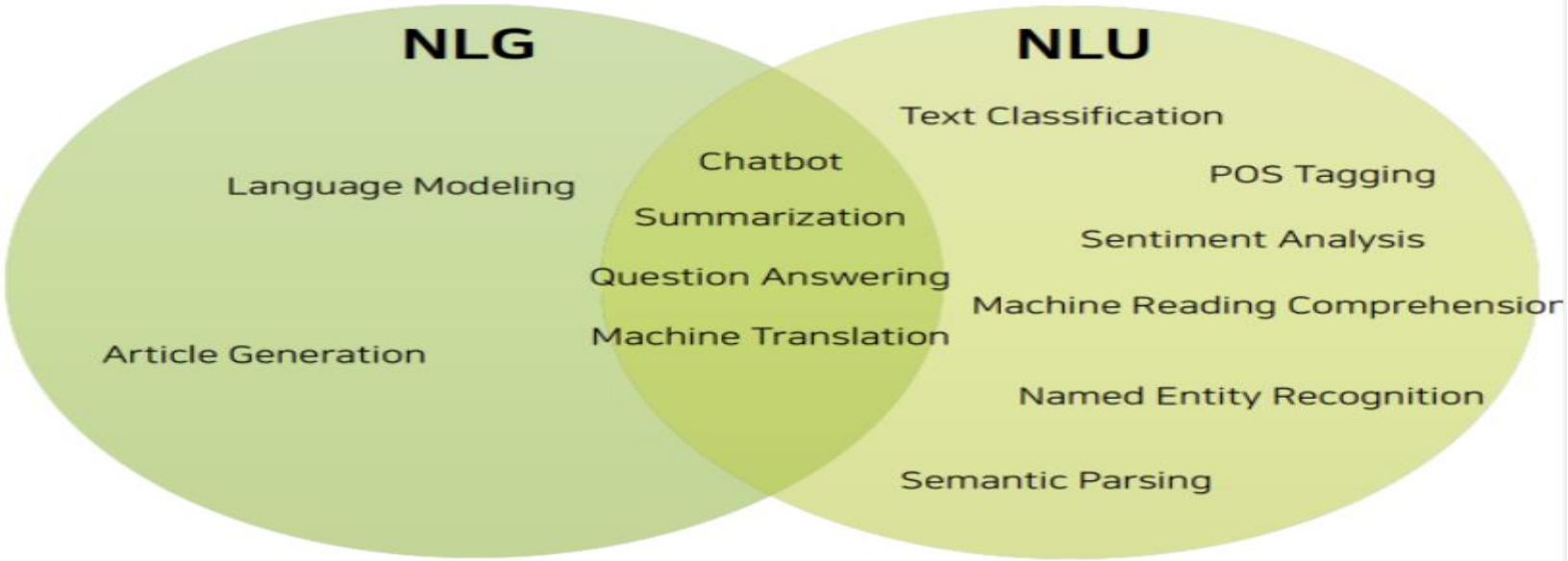
자연어? 인공어?





자연어처리란?

- 자연어처리란?
  - 사람들이 사용하는 일상적인 언어(자연어)를 컴퓨터를 이용하여 이해하고 생성하도록 하는 것을 뜻한다.
  - NLU(자연어이해),NLG(자연어생성)단계로 구분 가능하다.
    - ✓ NLU는 컴퓨터가 이해할 수 있는 값으로 바꾸는 것, NLG는 컴퓨터가 만든 언어를 사람이 이해할 수 있도록 바꾼 것을 뜻한다.



난이도가 높은 자연어처리

- 난이도가 높은 자연어처리
  - 언어의 중의성 및 모호성이 존재한다.

원문	차를 마시러 공원에 가던 차 안에서 나는 그녀에게 차였다.
G*	I was kicking her in the car that went to the park for tea.
M*	I was a car to her, in the car I had a car and went to the park.
N*	I got dumped by her on the way to the park for tea.
K*	I was in the car going to the park for tea and I was in her car.
S*	I got dumped by her in the car that was going to the park for a cup of tea.

원문	나는 철수를 안 때렸다.
1	철수는 맞았지만, 때린 사람이 나는 아니다.
2	나는 누군가를 때렸지만, 그게 철수는 아니다.
3	나는 누군가를 때린 적도 없고, 철수도 맞은 적이 없다.

난이도가 높은 자연어처리

- 난이도가 높은 자연어처리
  - 규칙의 예외가 존재한다.
    - ✓ 닫다 ->닫이다 ->닫히다
    - ✓ 닫다 -> 닫이다
    - ✓ 영어 동사의 과거형에는 -ed가 붙지만 아닌 경우가 존재

난이도가 높은 자연어처리

- 난이도가 높은 자연어처리
  - 언어는 유연성과 확장성이 무한하다.
    - ✓ 새로운 언어가 태어나는 방법도 다양하다.(잼민이)
    - ✓ 단어와 소리의 개수는 유한 하지만 이를 조합하여 만들 수 있는 문장은 무한하다.

번호	문장 표현
1	여자가 김치를 어떤 남자에게 집어 던지고 있다.
2	여자가 어떤 남자에게 김치로 때리고 있다.
3	여자가 김치로 싸대기를 날리고 있다.
4	여자가 배추 김치 한 포기로 남자를 때리고 있다.
5	여자가 김치를 사용해 남자를 때리고 있다.
6	남자가 여자에게 김치로 싸대기를 맞고 있다.
7	남자가 여자로부터 김치로 맞고 있다.

자연어처리 연구의 패러다임

- 자연어처리 연구의 패러다임
  - 규칙 기반
    - ✓ 언어의 문법적 규칙을 사전에 정의 하고 자연어를 처리한다.

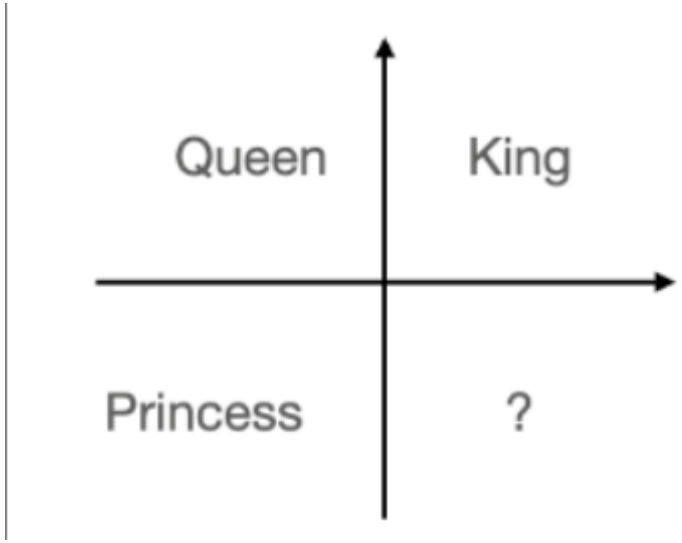
- 규칙 기반 자연어처리의 한계점
  - 어순이 정형화 되어 있지 않으면 한계가 크다.
  - 규칙을 미리 지정하는 것의 부담이 매우 크다.

자연어처리 연구의 패러다임

- 자연어처리 연구의 패러다임
  - 통계 기반
    - ✓ 언어의 규칙이 있다면 통계적으로 유의미한 값을 도출한다.
    - ✓ 어떤 단어가 선행 했을 때, 다른 단어가 나올 확률을 조건부 확률을 통해 계산한다.
- 통계 기반 자연어처리의 한계점
  - 사람의 손길이 많이 가는 방법이다.
  - 복잡한 규칙을 처리하기엔 어려움이 존재한다.

자연어처리 연구의 패러다임

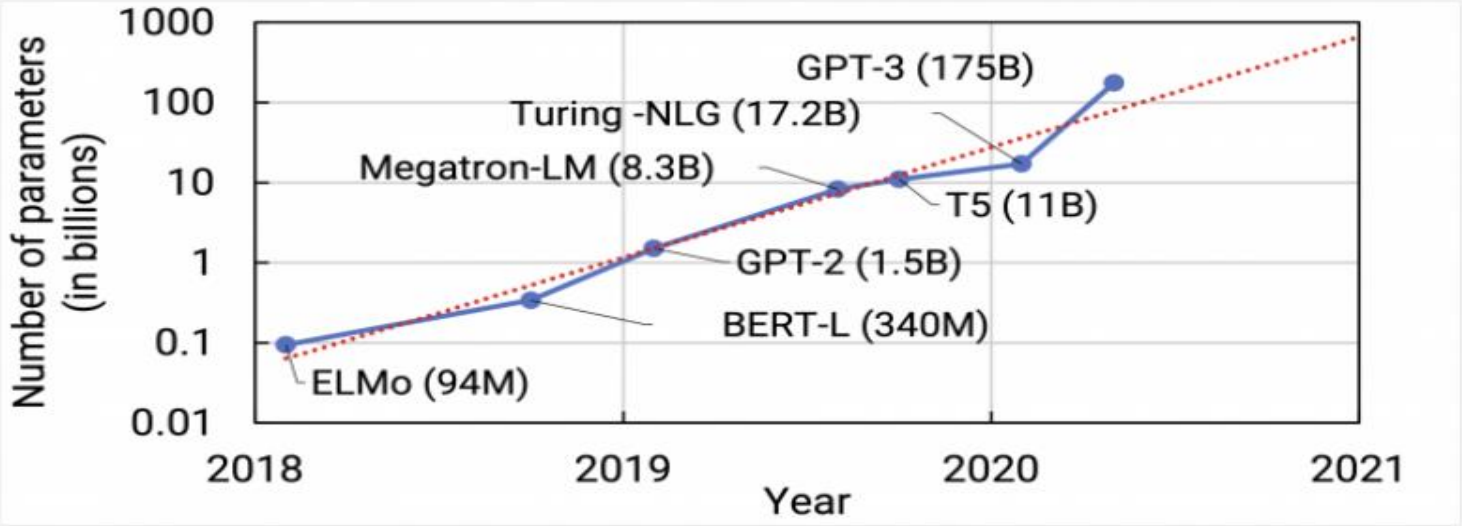
- 자연어처리 연구의 패러다임
  - 딥러닝 기반
    - ✓ 단어 임베딩
      - ❖ 단어를 벡터화 시켰다는 것인데, 단어의 주변을 보면 단어 파악이 가능 하다는 것이다.



자연어처리 연구의 패러다임

- 자연어처리 연구의 패러다임
  - 딥러닝 기반
    - ✓ 코퍼스
    - ✓ '말뭉치' 라고도 불리며 여러 단어들로 이루어진 문장을 뜻한다.

    코퍼스가 많고 오류가 적을수록 NLP모델은 정교해지고 정확도가 높아진다.

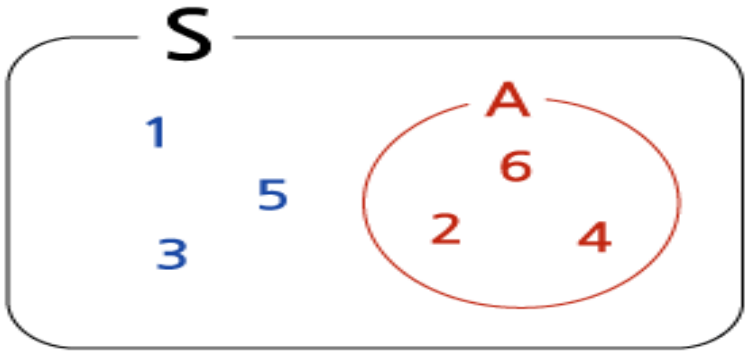


02

자연어 처리를 위한 수학

확률의 기초

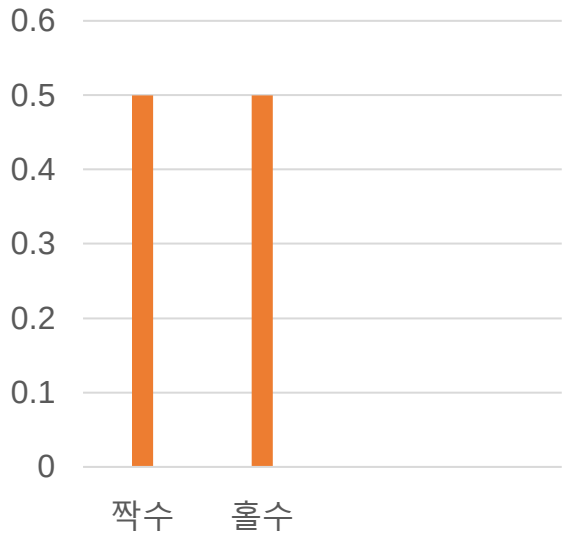
- 확률의 기초
  - 사건과 표본공간
    - ✓ 실험으로 나온 모든 결과를 담고 있는 집합은 표본공간이라 한다.
    - ✓ 사건은 표본공간의 부분집합 이다.
- 확률 변수
  - 시행의 결과에 따라 값이 결정되는 변수를 뜻한다.
  - 이산적인 사건을 표현하는 이산 확률 변수와 연속적인 사건을 표현하는 연속 확률 변수가 존재한다.

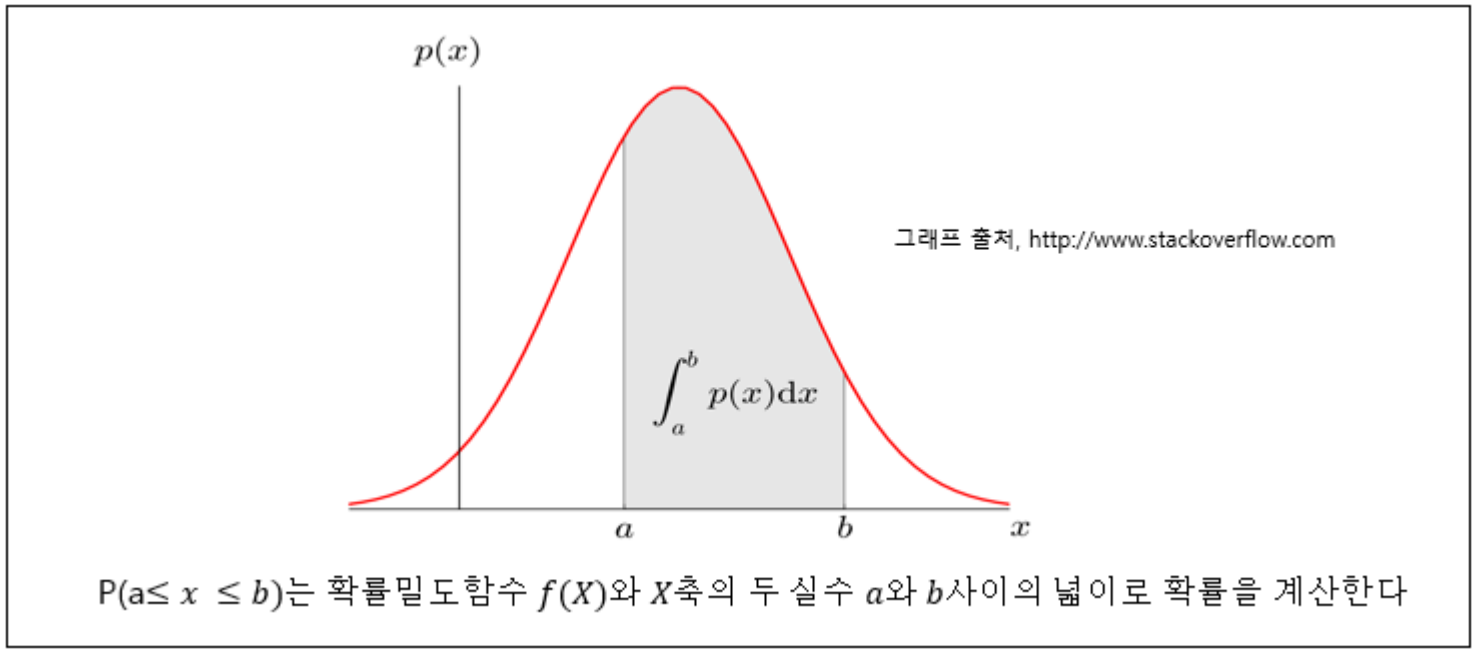


$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

확률의 기초

- 확률의 기초
  - 확률 분포
  - 확률 변수가 취할 수 있는 모든 값과 그 값들이 나타날 확률을 나열한 것 이다.

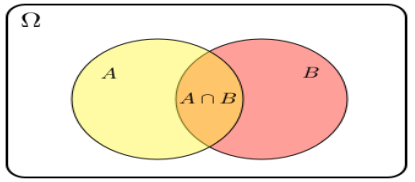




확률의 기초

- 확률의 기초
  - 조건부 확률
  - 어떤 사건 A가 일어났다고 가정한 상태에서 사건B가 일어날 확률을 의미

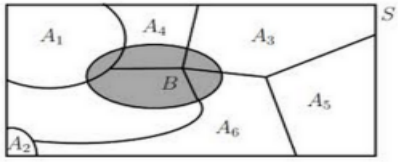
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$



- 베이즈 정리
  - 데이터라는 조건이 주어졌을 때의 조건부확률을 구하는 공식이다.

$$P(B) = \sum_{i=1}^{n} P(\mathrm{A}_i)P(B|A_i)$$

$$\mathrm{P}(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(\mathrm{A}_i)P(B|A_i)}$$



확률의 기초

- 확률의 기초
  - 기댓값과 분산 표준편차
    - ✓ 기댓값은 일종의 평균이라 볼 수 있다.
    - ✓ 분산은 확률 분포에서 확률 변수들의 퍼져있는 정도를 나타낸다.
    - ✓ 표준편차는 분산의 제곱근이다.

확률의 기초

- 확률의 기초
  - 이항분포
    - ✓ 확률이P인 베르누이 시행을 N번 반복시행할 때 출현 횟수를 나타내는 X의 분포를 의미한다. 이항분포의 표현은 B(n,p)로 하며 평균은np, 분산은 np(1- p)이다.
  - 다항분포
    - ✓ 이항분포의 일반화 이다.
    - ✓ 범주가 k개 이다

$$f(k;n,p) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

$$f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \frac{n!}{x_1!\ \ldots\ x_k!}p_1^{x_1}\ldots p_k^{x_k}$$

  - 정규분포
    - ✓ 연속 확률분포 중 하나이고, 자료의 분포를 근사하는데 자주 사용한다.

정규분포의 표현은 $N(\mu,\sigma^2)$로 하고 평균과 분산은 각각 μ, , σ^2 로 나타낸다.

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

MLE와 MAP

- MLE와 MAP
  - 각각 최대 우도 추정, 최대 사후 확률 추정을 의미한다.
  - MLE와 MAP는 수식적으로 연관성이 높고, MLE는 관측치에 영향을 많이 받으며 MLE는 MAP의 특수한 형태이다.

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}} P(h|D) \\
&= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)
\end{aligned}
$$

$$
h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}} P(D|h)
$$

정보이론과 엔트로피

- 정보이론과 엔트로피
  - 정보이론의 두가지
    - ✓ 중요성과 가법성
      - ❖ 중요성이란 어떤 사건이 발생할 가능성이 적을수록 그 사건은 많은 정보를 지닌다는 것 이다.

$$P(x_1) > P(x_2) \Rightarrow I(x_1) < I(x_2)$$

      - ❖ 가법성이란 어떤 두 사건x1,x2가 독립 이라면 다음을 만족 한다는 것 이다.

$$I(x_1 x_2) = I(x_1) + I(x_2)$$

정보이론과 엔트로피

- 정보이론과 엔트로피
  - 정보량
    - ✓ 앞서 말한 내용을 통해 정보량을 밑과 같이 나타낼 수 있다

$$I(x) = \frac{1}{P(x)}$$

- ✓ .정보량은 항상 양수 이다. $I(E) \geq 0$
- ✓ 따라서 정보량은 log함수로도 표현이 가능하다.(밑 수식은 통계적으로 독립일때를 뜻합니다.)

$$I(E_i E_k) = I(E_i) + I(E_k) = \log_2 (1/P_i) + \log_2 (1/P_k)$$

정보이론과 엔트로피

- 정보이론과 엔트로피
  - 엔트로피
  - 엔트로피란 확률변수 X의 표본공간에서 나타나는 모든 사상들의 정보량의 평균적인 기댓값을 의미한다.

$$H(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\log_2 P(X)]$$
$$= -\sum_x P(X = x) \log_2 P(X = x)$$

- KL-Divergence
  - 두 분포의 일치 정도를 측정하는 수단이다. Relative Entropy라고도 한다.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \qquad D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

$$D_{\mathrm{KL}}(P\|Q) = H(P,Q) - H(P)$$
$$= (-\sum p(x) \log q(x)) - (-\sum p(x) \log p(x))$$

03

언어학의 기본 원리

음절,형태소,어절,품사

- 음절,형태소,어절,품사
  - 음절
    - ✓ 언어를 말하고 들을 때, 한 덩어리로 여겨지는 가장 작은 단위를 말한다.

'이 문장에서 음절은 몇 개일까?' (1)

/이 문장에서 음저른 멷 개일까?/ (2)

  - 형태소
    - ✓ 언어에서 의미를 가지는 가장 작은 단위를 말한다.
    - ✓ 실질적인 의미의 유무에 따라 실질 형태소와 형식 형태소로 나뉜다.
    - ✓ 자립성의 유무에 따라 자립 형태소와 의존 형태소로 나뉜다.

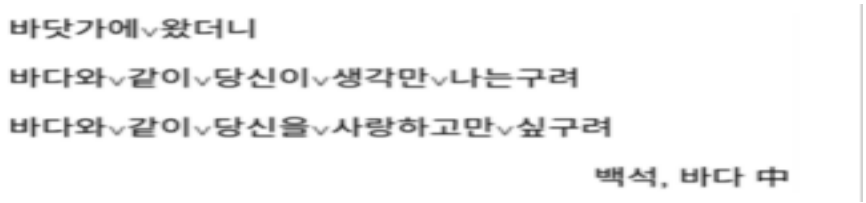
나는 컴퓨터 공부가 좋아.

실질 형태소 : '나', '컴퓨터', '공부', '좋-'
형식 형태소 : '는', '가', '아'

자립 형태소 : '나', '컴퓨터', '공부'
의존 형태소 : '는', '가', '좋-', '-아'

음절,형태소,어절,품사

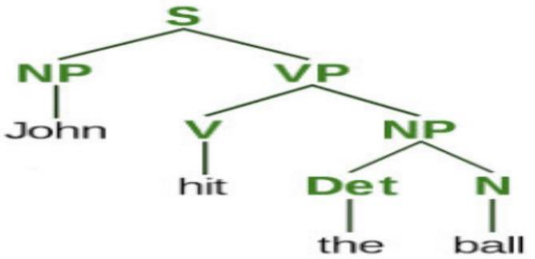
- 음절,형태소,어절,품사
  - 어절
    - ✓ 한 개 이상의 형태소가 모여 구성된 단위로 발화 시 어절을 중심으로 끊어서 말하고, 글을 쓸 때 에는 어절 단위로 띄어쓰기 한다.



  - 품사
    - ✓ 문장 내에서 단어의 역할을 기준으로 체언, 수식언,관계언,독립언, 용언으로 나눈다.
    - ✓ 형태에 따라서는 9품사(명사,대명사,수사,관형사,부사,조사,감탄사,동사,형용사) 로 나눈다.

구구조와 의존구조

- 구구조와 의존구조
  - 구구조
    - ✓ 구구조란 문장의 요소들이 서로 짝을 지어 구와 절을 이룸으로써 형성되는 구조이다.
    - ✓ 표면적으로 같은 언어요소와 순서로 되어 있어도 뜻이 다른 문장을 다른 구조로 기술하여 뜻을 파악하기 용이하다.

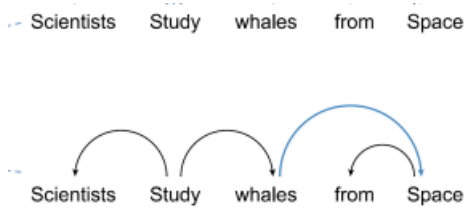


  - 의존구조
    - ✓ 단어들이 서로 의존관계를 이루어 하나의 구문을 형성한다고 보는 구조이다.
    - ✓ 구의 배열이 영어에 비해 자유로운 한국어 분석에 적절하다.

'Scientists study whales from space.'

1) '과학자들은 우주에서 고래에 대해 공부한다.'

2) '과학자들은 우주에 있는 고래에 대해 공부한다.'



의미론과 화용론

- 의미론과 화용론
  - 의미론
    - ✓ 문법적으로 옳은 문장이라도 의미가 어색하다면 언어라고 할 수 없다.

      예)'사료가 개를 먹었습니다'
  - 화용론
    - ✓ 언어 사용자와 발화 맥락을 고려하는 것을 의미한다.
    - ✓ 문맥을 이해하기 위해서는 상대방이 표현하는 것이 무엇인지를 알아야 한다.

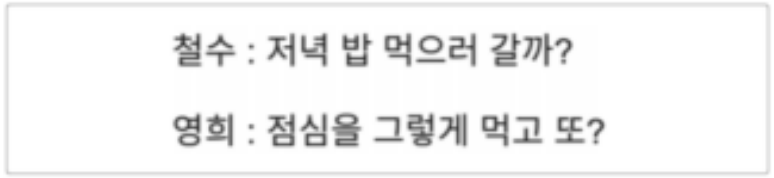


그림 3-11 화용론