Essential Information Theory

-based on Manning and Schutze – Foundations of Statistical Natural Language Processing-

Introduction

- **Information Theory**
  - 1948 년 Claude Shannon 이 처음 제안
  - 임의의 '정보'와 '통신채널'의 소스에 대해서 데이터 압축률과 전송률을 최대화시킬 수 있는 수학 모델 제시
  - 데이터 압축률(Data Compression) – **Entropy *H***
  - 전송률(Translation Rate) – **Channel Capacity *C***

Entropy (1)

- 정의

Let p(x) be the probability mass function of a random variable X, over a discrete set of symbols(or alphabet) X:

p(x) = P(X = x), x ∈ X

Entropy

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

$$= \sum_{x \in X} p(x) \log \frac{1}{p(x)}$$

$$= -E(\log p(x)) = E\left(\log \frac{1}{p(x)}\right)$$

Entropy's Property (1)

❑ 성질 **1 (Self-information)**

- the average uncertainty of a single random variable
  - 단일 확률 변수에 대하여 우리가 모르고 있는 정도 측정
  - $H(X) \geq 0$: 이 값이 클수록 예측이 어려우므로 제공되는 정보의 가치가 높아진다.
  - $H(X) = 0$ 인 경우는 X 가 완전히 결정되어 100% 예측이 가능하므로 새로운 정보를 제공할 필요가 없다.

❑ **Weighted Coin** 예제

- 앞면이 나올 확률과 뒷면이 나올 확률이 같은 동전의 경우

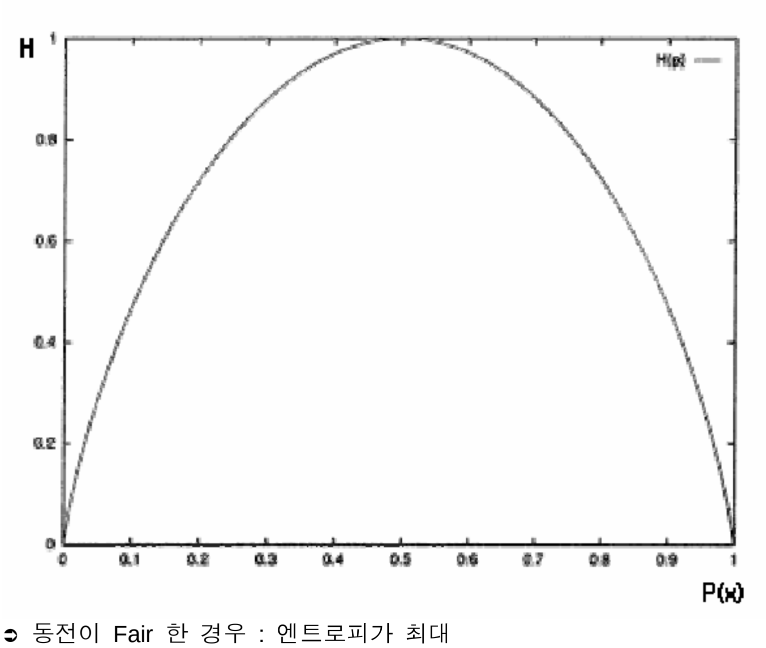
$$H\left(\frac{1}{2},\frac{1}{2}\right) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1\,bit$$

- 앞면이 나올 확률이 99% 인 동전의 경우

$$H\left(\frac{99}{100},\frac{1}{100}\right) = -\frac{99}{100}\log_2\frac{99}{100} - \frac{1}{100}\log_2\frac{1}{100} = 0.08\,bit$$

Entropy's Property (1) – Conti.

❑ **The entropy of weighted coin**



- 동전이 Fair 한 경우 : 엔트로피가 최대
- 동전이 Fair 하지 않은 경우 : 엔트로피가 작아짐

Entropy's Property (2)

❑ 성질 **2 (Entropy and number of bits)**

- the amount of information in a random variable

- the average length of the message needed to transmit an outcome of a random variable
  - 엔트로피는 메시지를 인코딩하기 위해 필요한 평균 비트 수

- 8-sided die 예제
  - 엔트로피

$$H(X) = -\sum_{i=1}^{8} p(i)\log p(i) = -\sum_{i=1}^{8} \frac{1}{8}\log\frac{1}{8} = -\log\frac{1}{8} = 3\,bits$$

  - 3 비트의 binary 메시지로 인코딩한 예

1	2	3	4	5	6	7	8
001	010	011	100	101	110	111	000

Entropy's Property (2) – Conti.

- Simplified Polynesian 예제
  - the letter frequency

p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8

  - per-letter entropy

$$H(p) = -\sum_{i \in \{p,t,k,a,i,u\}} p(i)\log p(i)$$

$$= -\left[4 \times \frac{1}{8}\log\frac{1}{8} + 2 \times \frac{1}{4}\log\frac{1}{4}\right] \quad = \quad 2\frac{1}{2}\,bits$$

  - design a code

p	t	k	a	i	u
100	00	101	01	110	111

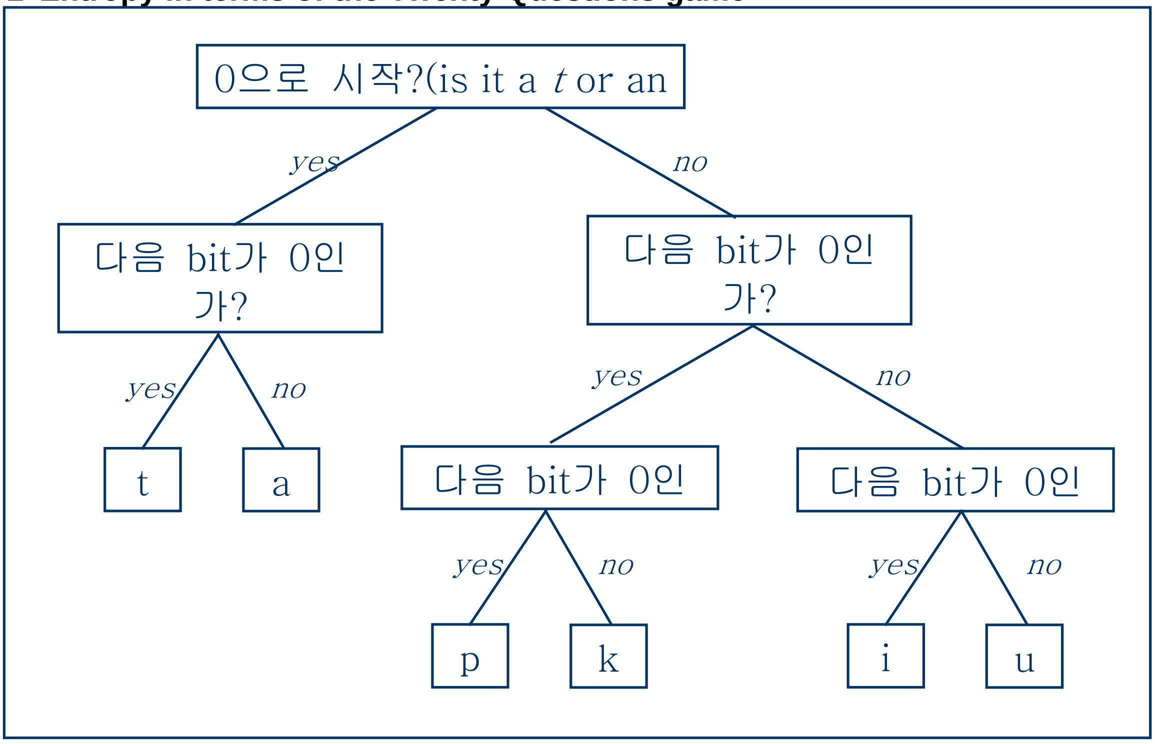
Entropy's Property (3)

- 성질 3 (Entropy and Search space)
  - the lower bound for the average number of bits needed to transmit that message
    - 비트를 많이 사용할수록 메시지가 무엇인지 알기 어렵다.
    - 이보다 적은 전송 비용을 들여 각 결과를 인코딩할 수 있는 더 좋은 방법은 없다.
  - a measure of the size of the 'search space'
    - 확률변수와 연관된 확률을 통하여 좋은 분류 기준을 선택할 수 있는 척도가 된다.
    - Twenty Question Game

– 엔트로피가 낮은(빈도수가 큰, 예측이 쉬운) 순서로 질문하는 것이 유리
– Simplified Polynesian 예제의 경우 각 문자를 확인하는데는 2 와 1/2 개의 질문만으로 충분

Entropy's Property (3) - Cont'd

❑ **Entropy in terms of the Twenty Questions game**



Joint entropy & Conditional entropy (1)

❑ **Joint Entropy**

discrete random variables X, Y ~ p(x,y)

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(X,Y)$$

❑ **Conditional Entropy**

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in X} p(x) H(Y|X=x) \\
&= \sum_{x\in X} p(x) \left[-\sum_{y\in Y} p(y|x) \log p(y|x) \right] \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y) \log p(y|x)
\end{aligned}
$$

Joint entropy & Conditional entropy (2)

- **Chain rule for entropy**

$$
H(X,Y) = H(X) + H(Y|X)
$$

$$
H(X_1,\ldots,X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1,\ldots X_{n-1})
$$

Proof)

$$
\begin{aligned}
H(X,Y) &= -E_{p(x,y)}(\log p(x,y)) \\
&= -E_{p(x,y)}(\log p(x)p(y|x)) \\
&= -E_{p(x,y)}(\log p(x) + \log p(y|x)) \\
&= -E_{p(x)}(\log p(x)) - E_{p(x,y)}(\log p(y|x)) \\
&= H(X) + H(Y|X)
\end{aligned}
$$

Joint entropy & Conditional entropy (3)

- **Simplified Polynesian revisited**
  - Simplified Polynesian has syllable structure
    - all words consist of sequences of CV(consonant-vowel)
  - C,V ~ p(c,v) 를 따른다고 할때, joint distribution

	p	t	k	P(.,V)
a	1/16	3/8	1/16	1/2
I	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
P(C,.)	1/8	3/4	1/16	

- the probabilities of the letters on a per-letter

p	t	k	a	I	u
1/16	3/8	1/16	1/4	1/8	1/8

Joint entropy & Conditional entropy (3) -Conti

- **Simplified Polynesian revisited**
  - the entropy of the joint distribution(by the chain rule)

$$H(C) = 2\times\frac{1}{8}\times 3+\frac{3}{4}(2-\log 3)$$
$$= \frac{9}{4}-\frac{3}{4}\log 3\, bits \approx 1.061\, bits$$

$$H(V\mid C) = \sum_{c=p,t,k} p(C=c)H(V\mid C=c)$$
$$= \frac{1}{8}H\left(\frac{1}{2},\frac{1}{2},0\right)+\frac{3}{4}H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)+\frac{1}{8}H\left(\frac{1}{2},0,\frac{1}{2}\right)$$
$$= 2\times\frac{1}{8}\times 1+\frac{3}{4}\left[\frac{1}{2}\times 1+2\times\frac{1}{4}\times 2\right]$$
$$= \frac{1}{4}+\frac{3}{4}\times\frac{3}{2} = \frac{11}{8}\, bits = 1.375\, bits$$

$$H(C,V) = H(C) + H(V\mid C)$$
$$= \frac{9}{4}-\frac{3}{4}\log 3+\frac{11}{8} = \frac{29}{8}-\frac{3}{4}\log 3 \approx 2.44\, bits$$

Joint entropy & Conditional entropy (4)

- **Entropy rate**
  - For message length n, the per-letter/word entropy

$$H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$$

  - The entropy of a human language L
    - a language is a stochastic process consisting of a sequence of tokens L = X(I)
    - the entropy rate for that stochastic process

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

Mutual Information (1)

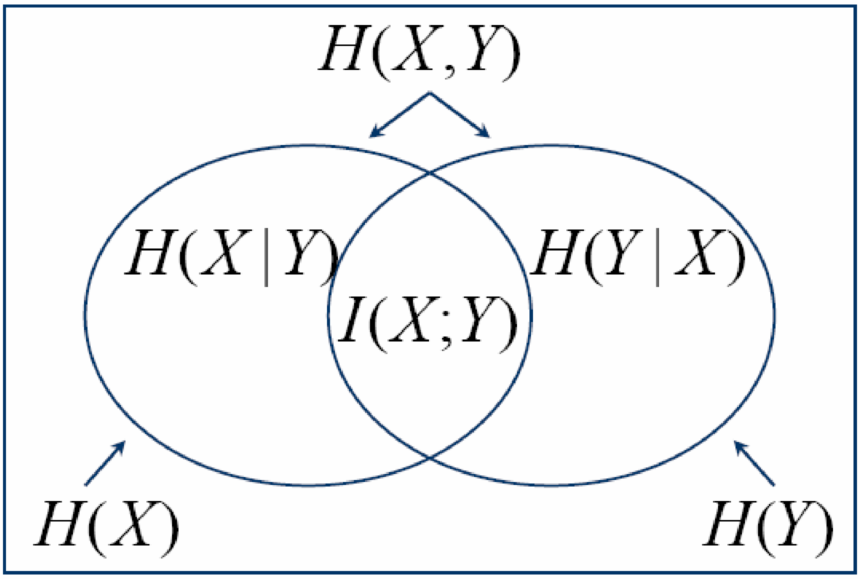
- **Mutual Information이란?**

$$\textit{By the chain rule for entropy},$$
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$
$$\textit{Therefore},$$
$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

  - the reduction in uncertainty of one random variable due to knowing about another
  - the amount of information one random variable contains about another
  - symmetric, non-negative measure of the common information in the two variables
  - a measure of dependance(or indepencdance) between variables

Mutual Information (2)

- **The relationship between MI(I) and Entropy(H)**



- ➲ It is 0 only when two variables are independent
- ➲ For two dependent variables, mutual information grows with
  - the degree of dependence
  - the entropy of the variables

Mutual Information (3)

- **Formulars for MI**

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= \sum_x p(x)\log\frac{1}{p(x)} + \sum_y p(y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p(x,y) \\
&= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)}
\end{aligned}
$$

$$\textit{Since } H(X \mid X) = 0, \textit{ note that}:$$

$$H(X) = H(X) - H(X \mid X) = I(X;X)$$

- ➲ conditional MI and chain rule

$$I(X;Y \mid Z) = I((X;Y) \mid Z) = H(X \mid Z) - H(X \mid Y,Z)$$

$$I(X_{1n};Y) = I(X_1;Y) + \ldots + I(X_n;Y \mid X_1,\ldots,X_{n-1})$$

$$= \sum_{i=1}^{n} I(X_i;Y \mid X_1,\ldots,X_{i-1})$$

Mutual Information (4)

- **Pointwise MI**
  - MI between two particular points in those distribution

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

  - a measure of association between elements
    - but there are problems with using this measure(see section 5.4)
- **MI 의 응용**
  - Clustering words(
  - Turn up in word sense disambiguation

Relative Entropy

- **Relative Entorpy(or Kullback-Leibler divergence)**

For two probability mass functions, p(x), q(x)

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p\left(\log \frac{p(x)}{q(x)}\right)$$

- the average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q
- D(p||q) ≥ 0 (D(p||q) = 0 인 경우 p = q)
- KL divergence(or KL distance) is not a metric
  - not symmetric in p and q
  - does not satisfy the triangle inequality

Cross Entropy (1)

- **Language Model vs. Entropy**
  - 언어 구조를 더욱 많이 포착할 수 있으려면, 모델의 엔트로피가 더욱 작아져야 한다.
  - 즉, 엔트로피는 모델의 품질을 측정할 수 있는 수단
- **Pointwise Entropy**
  - a matter of how surprised we will be

$$H(w \mid h) = -\log_2 m(w \mid h)$$

  - w : next word
  - h : history of words seen so far
- Total suprise

$$H_{total} = -\sum_{j=1}^{n} \log_2 m(w_j \mid w_1, w_2, \ldots, w_{j-1}) = -\log_2 m(w_1, w_2, \ldots, w_n)$$

Cross Entropy (2)

- **Real Distribution vs. Model**
  - 일반적으로, 발화(utterance) 와 같은 경험적 현상이 어떤 확률 분포(p)를 가질 것인지를 알 수 없다.
  - 그러나, 발화의 코퍼스(corpus)를 관찰하여 대략적으로 확률 분포(m)를 추정할 수 있다.
  - 이렇게 추정한 확률 분포(m)를 실제 확률 분포(p)에 대한 모델이라고 한다.

- **Model 의 요건**
  - minimize D(p||m)
    - 가능한 확률적으로 정확한 분포를 모델로 결정
    - but, do not know what p is.
    - cross entropy 개념을 이용하여 이를 해결

Cross Entropy (3)

- **Cross Entropy의 정의**

X ~ true probablility distribtion p(x)

q : normally a model of p

$$H(X,q) = H(X) + D(p \| q) = -\sum_x p(x) \log q(x)$$

$$= E_p\left(\log \frac{1}{q(x)} \right)$$

- **Cross Entropy of Language**

$$H(L,m) = -\lim_{n \to \infty} \frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log m(x_{1n})$$

$$= -\lim_{n \to \infty} \frac{1}{n} \log m(x_{1n})$$

$$\approx -\frac{1}{n} \log m(x_{1n})$$

$$H(L,m) = \lim_{n \to \infty} \frac{1}{n} E\left(\log \frac{1}{m(X_{1n})} \right)$$

Assumption that the language is 'nice'

For a sufficiently large n

Expectation embedded

Cross Entropy (4)

- **Asymptotic Equipartition Property**
  - Shannon-McMillan-Breiman theorem 의 결론
  - If H_{rate} is the entropy rate of a finate-valued stationary ergodic process (X_n), then:

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \rightarrow H_{rate} \ with \ probability \ 1$$

- ergodic process
  - cannot get into different substates that it will not escape from
  - 코퍼스를 장기간 관찰하여 얻은 확률분포는 실제 Language의 확률분포와 유사해진다.
- stationary process
  - do not change over time
  - 시간에 따라 통계적인 특성이 변하지 않음

- Language 는 시간에 따라 변하므로 엄밀히는 맞지 않지만, 일정 기간동안 언어는 변하지 않는다고 가정

The entropy of English (1)

- **Stochastic Models of English**
  - **n-gram models**
  - **Markov chains**
    - k^{th} order Markov approximation
      - Probability of the next word depends only on
        - the previous *k* words in the input(이전 k 개의 단어만 고려)

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_n = x_n \mid X_{n-1}, \ldots X_{n-k} = X_{n-k})$$

      - e.g. character basis
        - Guess what the next character in a text will be given the preceding *k* characters.

The entropy of English (2)

- **Assumption of simplified model of English**
  - The cross entropy gives us an upper bound for the true entropy of English.
  - since D(p||m) ≥ 0, H(X, m) ≥ H(X)

- **Model vs. Cross Entropy**

Model	Cross entropy(bits)
Zeroth order	4.76
First order	4.03
Second order	2.8
Shannon's experiment	1.3(1.34)

uniform model (=log 27)

Perplexity

- **People tend to refer to**
  - *Perplexity* rather than cross entropy
    - In the speech recognition community.

$$perplexity(X1n, m) = 2^{H(x1n,m)}$$
$$= m(x_{1n})^{-\frac{1}{n}}$$

- **Perplexity of *k***
  - You are as surprised on average as you would have been
    - If you had to guess between *k* equiprobable choices at each step.