ECE8813
Statistical Language Processing

Lecture 3: Information Theory Foundations

Chin-Hui Lee

School of Electrical and Computer Engineering

Georgia Institute of Technology

Atlanta, GA 30332, USA

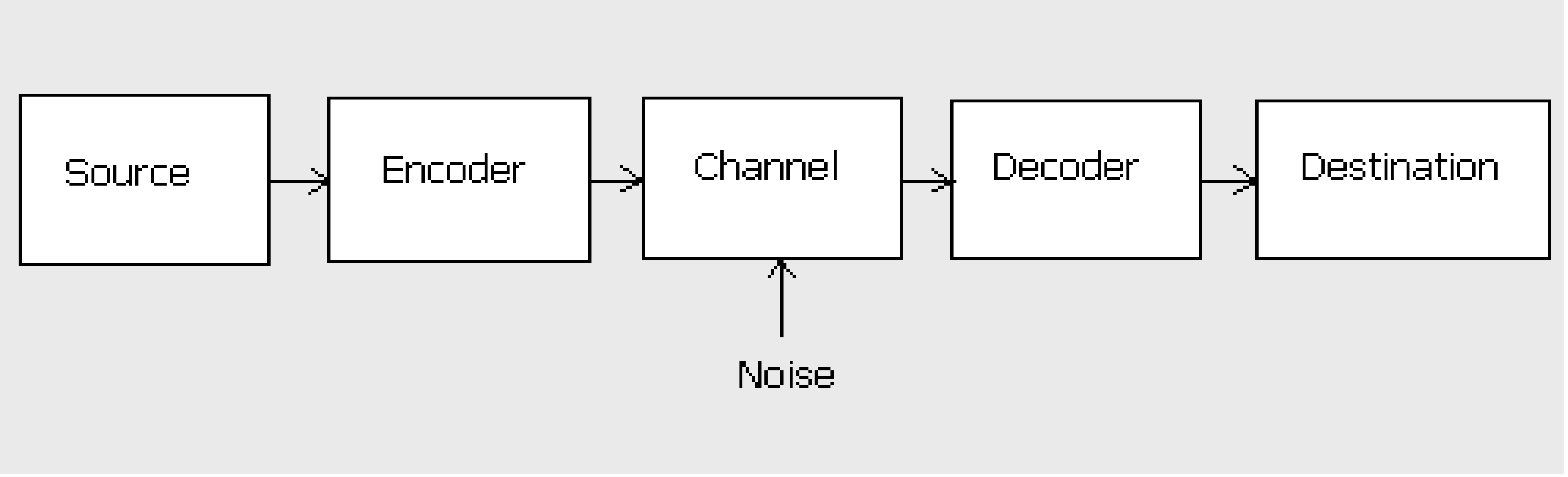
chl@ece.gatech.edu

Course Information

- Subject: Statistical Language Processing
- Prerequisite: ECE3075, ECE4270
- Background Expected
  - Basic Mathematics and Physics
  - Digital Signal Processing
  - Basic Discrete Math, Probability Theory and Linear Algebra
- Tools Expected:
  - MATLAB and other Programming Tools
  - Language-specific tools will be discussed in Class
- Teaching Philosophy
  - Textbooks and reading assignments: your main source of learning
  - Class Lectures: exploring beyond the textbooks
  - Homework: hand-on and get-your-hands-dirty exercises
  - Class Project: a good way to go deeper into a particular topic
- **Website:** http://users.ece.gatech.edu/~chl/ECE8813.sp09

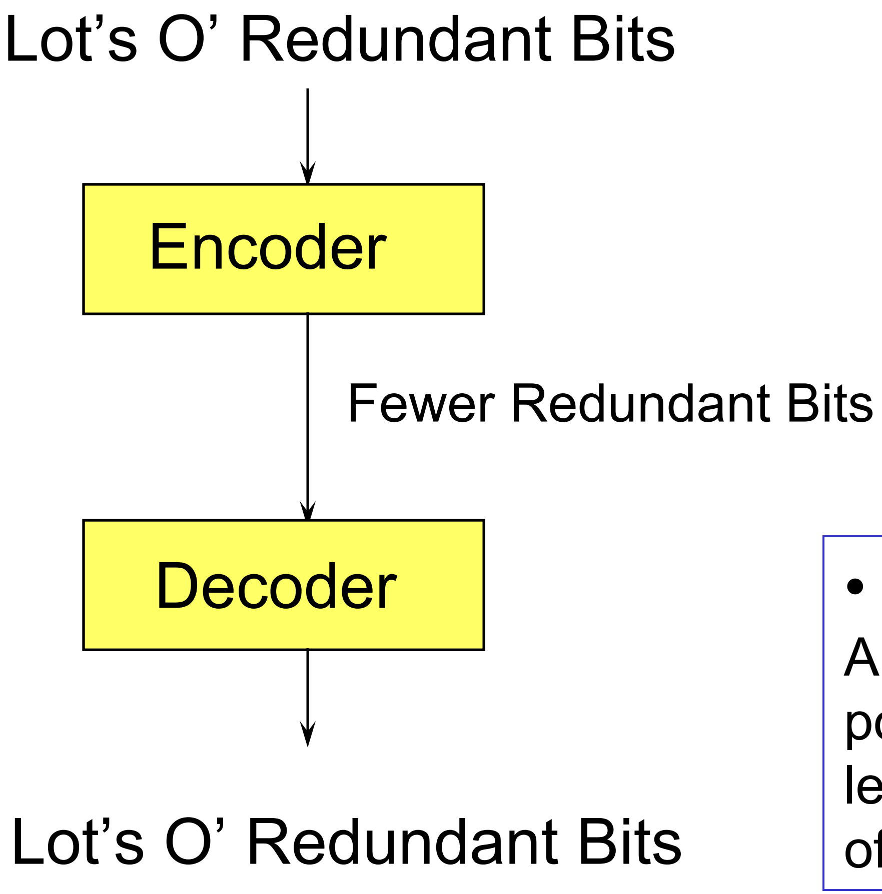
Information Theoretic Perspective

• Communication theory deals with systems for transmitting information from one point to another



• Information theory was born with the discovery of the fundamental laws of data compression and transmission, including channel modeling

Data Compression



- An interesting consequence:

A Data Stream containing the most possible information possible (i.e. the least redundancy) has the statistics of random noise

Huffman Coding

- Suppose we have an alphabet with four letters *A, B, C, D* with frequencies:

A	B	C	D
0.5	0.3	0.1	0.1

- Represent this with *A*=00, *B*=01, *C*=10, *D*=11. This would mean we use an average of 2 bits per letter
- On the other hand, we could use the following representation: *A*=1, *B*=01, *C*=001, *D*=000. Then the average number of bits per letter becomes

$$(0.5)*1+(0.3)*2+(0.1)*3+(0.1)*3 = 1.7$$

- The representation, on average, is more efficient.

Information Theory & C. E. Shannon

- Claude E. Shannon (1916-2001, from BL to MIT): Information Theory, Modern Communication Theory
- Entropy (Self-Information) – b*it,* amount of info in r.v.
- Study of English – Cryptography Theory, *Twenty Questions* game, Binary Tree and Entropy, etc.
- Concept of Code – Digital Communication, Switching and Digital Computation (optimal Boolean function realization with digital relays and switches)
- Channel Capacity – Source and Channel Encoding, Error-Free Transmission over Noisy Channel, etc.
- "A Mathematical Theory of Communication", Parts 1 & 2, *Bell System Technical Journal*, 1948.

Information vs. Physical Entropy

- Physicist Edwin T. Jaynes identified a direct connection between Shannon entropy and physical entropy in 1957
- Ludwig Boltzmann's grave is embossed with his equation: $S = k \log W$

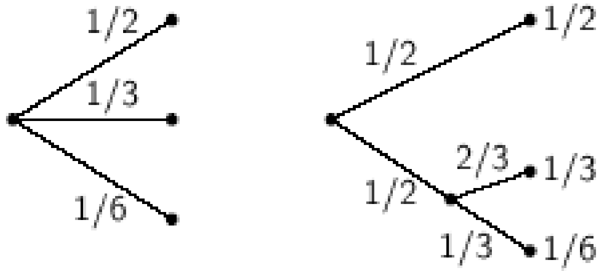
  Entropy = Boltzmann's-constant

  * log(function of # of possible micro-states)
- Shannon's measure of information (or uncertainty or entropy) can be written: $I = K \log \Omega$

Uncertainty

- Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$
- Say these probabilities are known, but that is all we know concerning which event will occur next
- What properties would a measure of our uncertainty, $H(p_1, p_2, \ldots, p_n)$, about the next symbol require:
  - — H should be continuous in the p_i
  - — If all the p_i are equal ($p_i = 1/n$), then H should be a monotonic increasing function of n
    - With equally likely events, there is more choice, or uncertainty, when there are more possible events
  - — If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Illustration on Uncertainty



- On the left, we have three possibilities:

  $p_1 = 1/2, p_2 = 1/3, p_3 = 1/6$

- On the right, we first choose between two possibilities:

  $p_1 = 1/2, p_2 = 1/2$

  and then on one path choose between two more:

  $p_3 = 2/3, p_4 = 1/3$

- Since the final probabilities are the same, we require:

  $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + 1/2\, H(2/3, 1/3)$

Entropy

- In a proof that explicitly depends on this decomposibility and on monotonicity, Shannon establishes

  Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

  where K is a positive constant

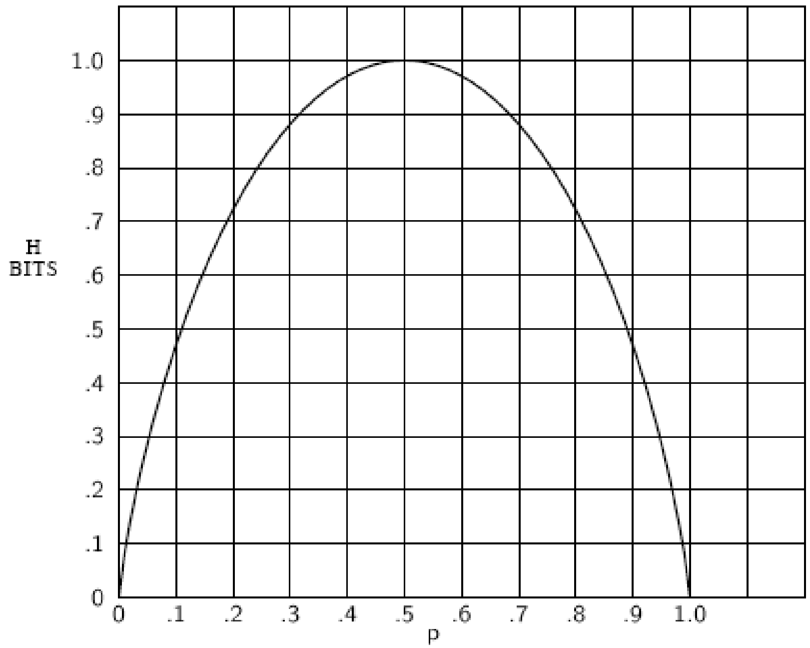
- Observing the similarity in form to entropy as defined in statistical mechanics, Shannon dubbed H the entropy of the set of probabilities $p_1, p_2, \ldots, p_n$
- Generally, the constant *K* is dropped; Shannon explains it merely amounts to a choice of unit of measure

Behavior of the Entropy Function

- In the simple case of two possibilities with probability p and $q = 1 - p$, entropy takes the form

$$H = -(p \log p + q \log q)$$

and is plotted here as a function of p:



More on the Entropy Function

- In general, $H = 0$ if and only if all the p_i are zero, except one which has a value of one
- For a given n, H is a maximum (and equal to log n) when all p_i are equal ($1/n$)
  - Intuitively, this is the most uncertain situation
- Any change toward equalization of the probabilities $p_1, p_2, \ldots, p_n$ increases H
  - If $p_i \neq p_j$, adjusting p_i and p_j so they are more nearly equal increases H
  - Any "averaging" operation on the p_i increases H

Joint Entropy

- For two events, *x* and *y*, with m possible states for *x* and *n* possible states for *y*, the entropy of the joint event may be written in terms of the joint probabilities

$$H(X,Y) = -\sum_{i,j} p(x_i,y_j) \log p(x_i,y_j)$$

while

$$H(X) = -\sum_{i,j} p(x_i,y_j) \log \sum_{j} p(x_i,y_j)$$

$$H(Y) = -\sum_{i,j} p(x_i,y_j) \log \sum_{i} p(x_i,y_j)$$

- It is "easily" shown that $H(X,Y) \leq H(X) + H(Y)$
  - Uncertainty of a joint event is less than or equal to the sum of the individual uncertainties
  - Only equal if the events are independent: $p(x,y) = p(x)\, p(y)$

Conditional Entropy

- Suppose there are two chance events, *x* and *y*, not necessarily independent. For any particular value x_i that *x* may take, there is a conditional probability that *y* will have the value y_j, which may be written

$$p(y_j|x_i) = p(x_i,y_j) / \sum_j p(x_i,y_j) = p(x_i,y_j) / p(x_i)$$

- Define the *conditional entropy* of *y*, *H(y|x)* as the average of the entropy of *y* for each value of *x*, weighted according to the probability of getting that particular *x*

$$H(Y|X) = -\sum_{i,j} p(x_i)\, p(y_j|x_i) \log p(y_j|x_i)$$

$$H(Y|X) = -\sum_{i,j} p(x_i,y_j) \log p(y_j|x_i)$$

  - This quantity measures, on the average, how uncertain we are about *y* when we know *x*

Joint, Conditional, & Marginal Entropy

- Substituting for $p(y_j|x_i)$, simplifying, and rearranging yields: $H(X,Y) = H(X) + H(Y|X)$
  - The uncertainty, or entropy, of the joint event *x*, *y* is the sum of the uncertainty of *x* plus the uncertainty of *y* when *x* is known
- Since $H(X,Y) \leq H(X) + H(Y)$, and given the above, then $H(Y) \geq H(Y|X)$
  - The uncertainty of *y* is never increased by knowledge of *x*
    - It will be increased unless *x* and *y* are independent, in which case it will remain unchanged

Conditioning Reduces Uncertainty

Interpretation: on the average, knowing about *Y* can only reduce the uncertainty about *X*

Y \ X	1	2
1	0	3/4
2	1/8	1/8

$$p(x) = \sum_y p(X,Y) \Rightarrow p(x=1) = \sum_y p(1,y) = \frac{1}{8}$$

$$p(x=2) = \sum_y p(2,y) = \frac{7}{8}$$

$$H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544 \text{ bits}$$

$$H(X|Y=1) = -\sum_x p(x|1)\log p(x|1) = 0 - \frac{3}{4}\log\frac{3}{4} = 0.3113$$

$$H(X|Y=2) = -\sum_x p(x|2)\log p(x|2) = -\frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{3}{4}$$

$$H(X|Y) = \frac{3}{4}H(X|Y=1) + \frac{1}{4}H(X|Y=2) = 0.4210$$

The uncertainty of *X* is decreased if *Y*=1 is observed, it is increased if *Y*=2 is observed, and is decreased on the average

Maximum and Normalized Entropy

- *Maximum entropy*, when all probabilities are equal is

  $H_{\text{max}} = \log n$

- Normalized entropy is the ratio of entropy to maximum entropy

  $H_o(X) = H(X) / H_{\text{max}}$

- Since entropy varies with the number of states, n, normalized entropy is a better way of comparing across systems
  - Shannon called this *relative entropy*
  - Some cardiologists and physiologists call entropy divided by total signal power normalized entropy

Mutual Information (MI)

- Define *Mutual Information* (aka *Shannon Information Rate*) as

$$I(X,Y) = \sum_{i,j} p(x_i,y_j) \log [\, p(x_i,y_j) \,/\, p(x_i)p(y_j) \,]$$

- When *x* and *y* are independent $p(x_i,y_j) = p(x_i)p(y_j)$, so $I(x,y)=0$
- When *x* and *y* are the same, the MI of *x*, *y* is the same as the information conveyed by *x* (or *y*) alone, which is just $H(x)$
- Mutual information can also be expressed as

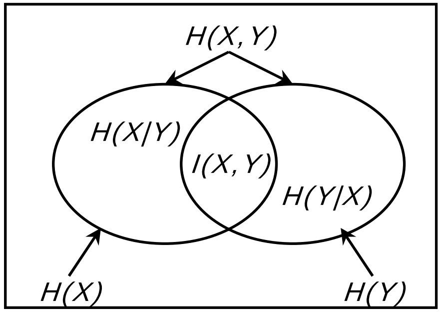
$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- Mutual information is nonnegative
- Mutual information is symmetric; i.e., $I(X,Y) = I(Y,X)$

Mutual Information

Definition :

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



$$I(X,Y) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} + \sum_{y \in Y} p(y) \log_2 \frac{1}{p(y)} - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{1}{p(x,y)}$$

- Show:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Point-wise Mutual Information

- Point-wise MI: the amount of information provided by the occurrence of the event represented by "*y*" about the occurrence of the event represented by "*x*"
- Event-specific not ensemble average

$$i(x, y) = \log_2 \frac{P(x \mid y)}{P(x)} = -\log_2 \frac{P(x)}{P(x \mid y)}$$

Entropy Definition Recap

- Entropy and information: given a discrete information source *x* with a pmf *p(x),* the number of bits required to describe the "information content" of the source

$$H(X) = -\sum_{x \in \mathrm{X}} p(x)\log_2 p(x) = \mathrm{E}[\log_2 \frac{1}{p(X)}] \quad 0\log_2 0 = 0$$

- Classical statistical thermodynamics

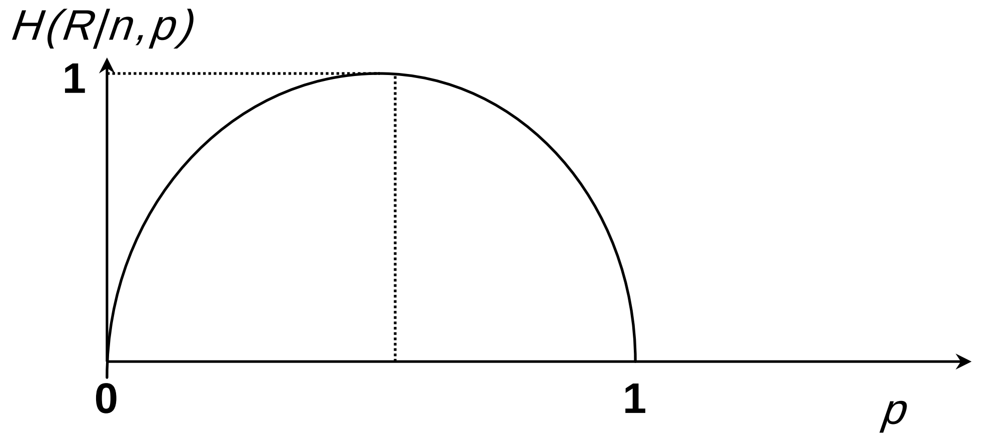
- Cross entropy and divergence

Entropy for Binomial Distributions

- Binomial distribution: Compute *H(R|n,p), n=1,2,…*

$$B(r;n,p)=\frac{n!}{r!(n-r)!}p^{r}(1-p)^{n-r} \quad \text{where} \quad 0\le r\le n$$

- Show *n=1, H(R|n,p)=1* peaks at *p=1/2 (worst case!)*



- How about for n=2 or more?
  - can you show max *H(R|n,p)=n* and peaks at *p=1/2* for all *n?*

Entropy Chain Rule

- Chain Rule for Entropy - Show the followings:

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

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

- Independence:

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

Conditional Mutual Information

- Conditional Mutual Information

$$I(X,Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

- Chain Rule for Mutual Information

$$I(X_1, X_2, \ldots, X_n, Y) = \sum_{i=1}^{n} I(X_i, Y | X_1, \ldots, X_{i-1})$$

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

Bayes' Theorem

- Swapping dependency between events
  - calculate P(B|A) in terms of P(A|B) that is available and more relevant in some cases $P(B|A)=\frac{P(B\cap A)}{P(A)}=\frac{P(A|B)P(B)}{P(A)}$
- In many cases, it is not important to compute P(A)

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

- Another Form of Bayes' Theorem (try n=2)
  - If a set B partitions A, i.e. $A=\bigcup_{i=1}^{n} B_i \quad B_i \cap B_k = \phi$

$$P(B_j|A)=\frac{P(A|B_j)P(B_j)}{P(A)}=\frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Kullback-Leibler (KL) Divergence

- Distance measure between pmf's (relative entropy)
  - *D(p||q)=0* if and only if *q=p*
  - Relative (cross) entropy between true *p(x)* and assumed *q(x)*

$$D(p \| q) = \mathrm{E}_p[\log_2 \frac{p(x)}{q(x)}] = \sum_{x \in \mathrm{X}} p(x) \log_2 \frac{p(x)}{q(x)}$$

- *KL Divergence* is a measure of the average number of bits that are wasted by encoding source *p(x)* with an estimated but not correct distribution *q(x)*
- Divergence can be a measure of independence, show that:

$$I(X,Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y))$$

Relative Entropy & Mutual Information

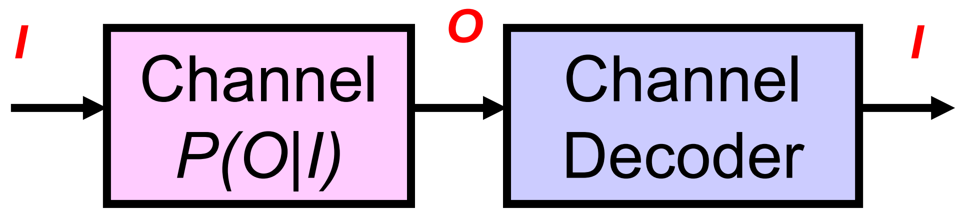
- Conditional Relative Entropy

$$D(p(x,y)\,||\,q(x,y)) = D(p(x)\,||\,q(x)) + D(p(y|x)\,||\,q(y|x))$$

- Chain Rule for Mutual Information

$$D(p(y|x)\,||\,q(y|x)) = \sum_{x\in \mathrm{X}} p(x) \sum_{y\in \mathrm{Y}} p(y|x) \log_2 \frac{p(y|x)}{q(y|x)}$$

Shannon's Channel Modeling Paradigm



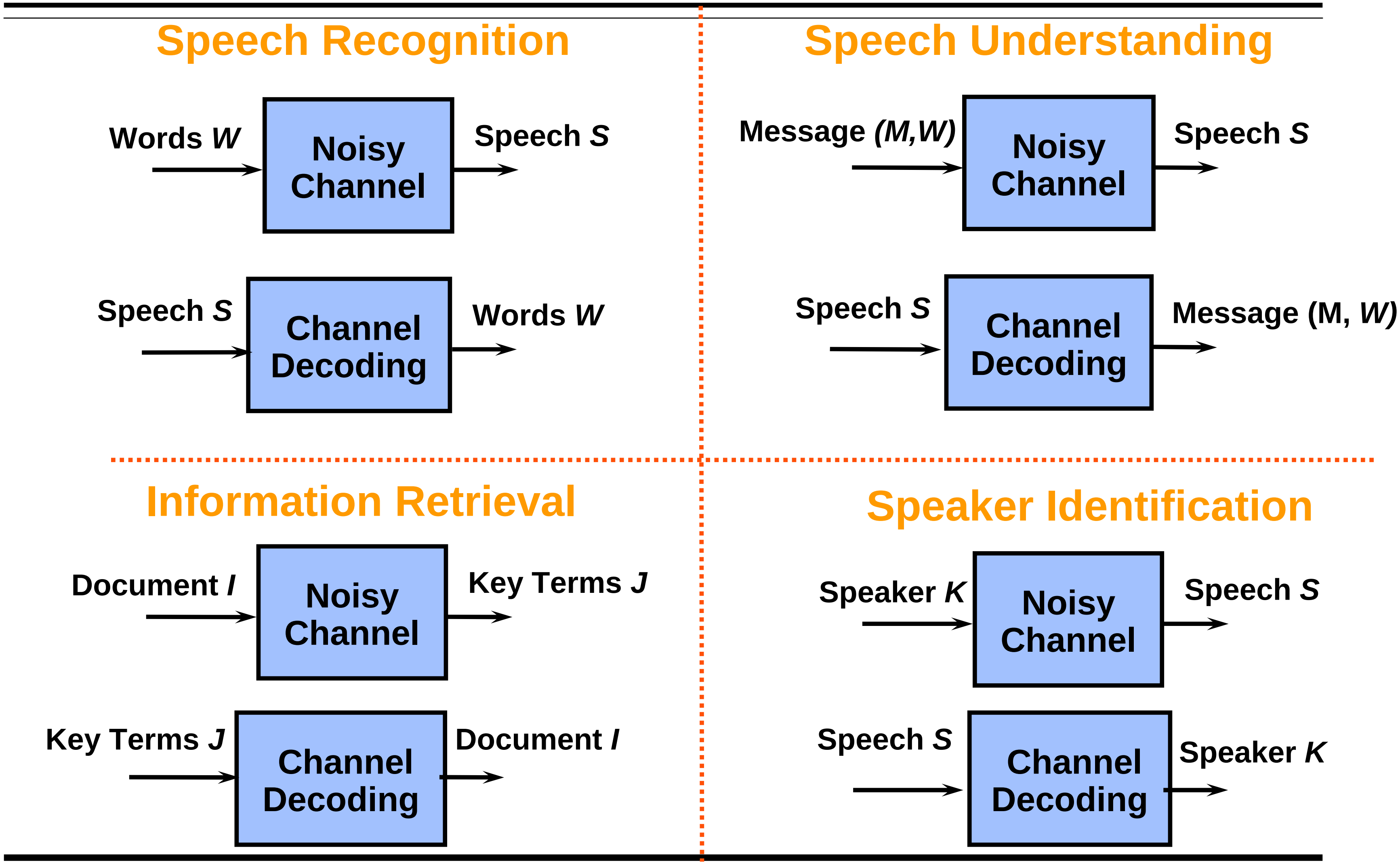
$$\hat{I} = \arg\max_{I \in \Omega} P(I \mid O) = \arg\max_{I \in \Omega} \frac{P(O \mid I)P(I)}{P(O)}$$

- Channel input is hidden (unobserved) while output is observed and used to infer the input (which is often approximated by a structural Markov model)
- Channel modeling with (*I*, *O*) pairs in large training sets

Modeling Input-Output Associations

- Hidden Markov Model (HMM)
- Artificial Neural Network (ANN)
- Classification and Regression Tree (CART)
- Support Vector Machine (SVM)
- Mixture of experts, Bayesian network
- Many New Applications
  - Rule induction, statistical parsing, machine translation
  - Information retrieval, text categorization, call routing, transliteration, pronunciation, machine translation, etc.

Channel Modeling and Decoding



Study on Entropy of English Letters

Model	Cross Entropy (bits)	Comments
Zeroth order	4.76	uniform letter log(27)
First order	4.03	unigram
Second order	2.8	bigram
Shannon's 2^{nd} Experiment	1.34	human prediction

Students' in-class computations verify results, and trigram ~ 2 bits

C. E. Shannon, "Prediction and Entropy of Printed English", *Bell System Technical Journal*, Vol. 30, pp. 50-64, 1951.

Probabilities of Letter Sequences

Markov Approximation to Probability of Letters

$$P(L) = P(l_1)P(l_2 \mid l_1)\cdots P(l_{|L|} \mid l_1,\ldots,l_{|L|-1}) \quad k-gram$$

$$\approx P(l_1)P(l_2 \mid l_1)\cdots P(l_k \mid l_1,\ldots,l_{k-1})\prod_{i=k+1}^{|L|} P(l_i \mid l_{i-1},l_{i-2},\ldots,l_k)$$

- Cross entropy between true *p(x)* and model *q(x)*

$$H(X,q) \equiv H(X) + D(p(x) \,\|\, q(x)) = -\sum_{x\in \mathrm{X}} p(x)\log_2 q(x) = \mathrm{E}_{\mathrm{p}}[\log_2 \frac{1}{q(X)}]$$

- Perplexity: branching factor

$$H(X) \approx \log_2(\mathrm{Perp}(X))$$

Entropy and Language Modeling

- Cryptography: the Enigma machine
  - Units and their co-occurrence statistics
  - Encryption and decryption of “fixed” units
  - Language ID of encrypted sources
- Information retrieval & text classification
  - Words as units and document modeling
- Multimedia pattern recognition
  - Definition and modeling of audiovisual alphabets
  - Tokenization: converting media to unit sequences
  - Representation of audiovisual patterns
  - Language modeling of units and co-occurrences
  - Discriminative classifier learning

Summary

- Today's Class
  - Information Theory Foundations
  - Web: http://www.ece.gatech.edu/~chl/ECE8813.sp09
- Next Class
  - Optimization essentials on Jan. 15
- Reading Assignments
  - Manning and Schutze, Chapters 1 & 2