

Tutorial on Information Theory I

Robert Daland
Assistant Professor
UCLA, Linguistics

What is 'information'?

- Logic/Philosophy/Semantics
 - the 'world' contains a number of entities and relations
 - rational perceiver: incomplete knowledge
 - 'information' is any message which reduces our uncertainty about the 'world'

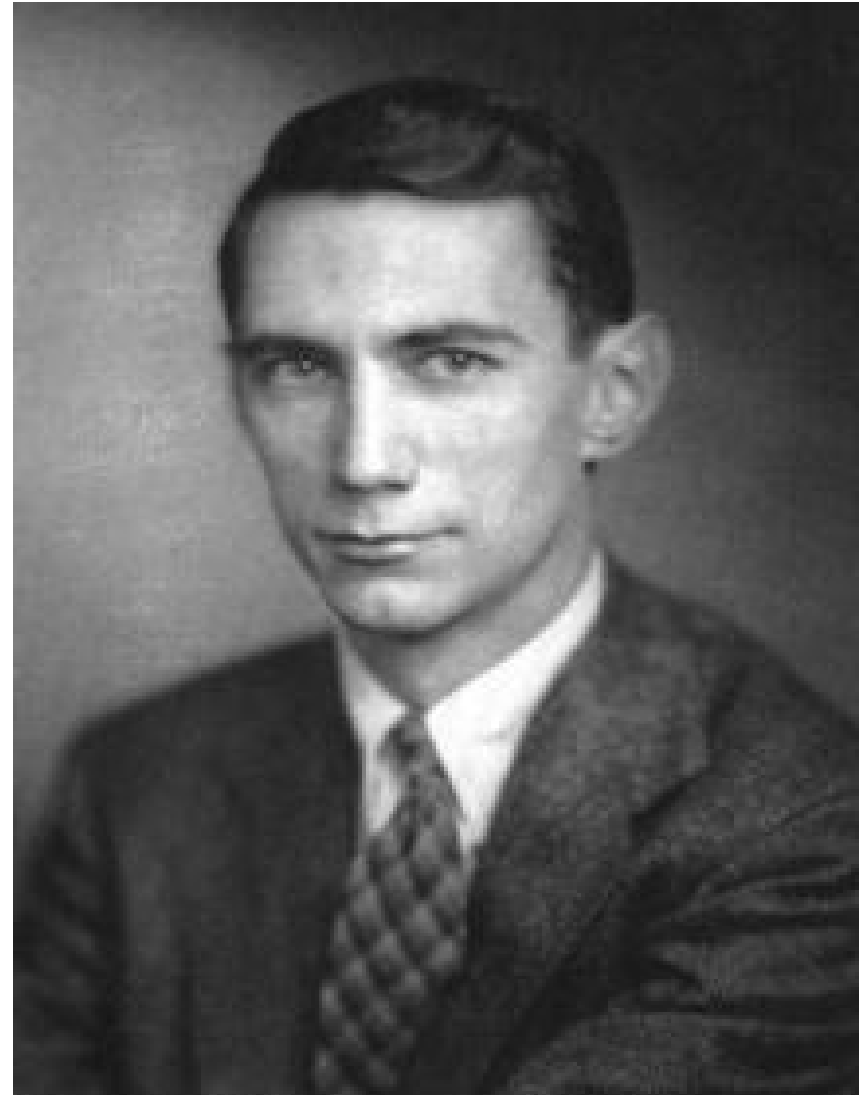
Structure of the talk

- Introduction [done!]
- History
- Foundation: Entropy
- Extensions: Conditional entropy, cross entropy
- Applications

History

Post-WWII

- WWII dramatically illustrated the need for (secured) long-distance communication
- two problems:
 - cost
 - noise



Noisy channel model



Codes

- A 'code' is a set of strings over an alphabet Σ
- Each such string is called a 'message'
- Typically, the alphabet is *binary*: $\Sigma = \{0, 1\}$



PAUL REVERE'S RIDE.

LISTEN, my children, and you shall hear
Of the midnight ride of Paul Revere,
On the eighteenth of April, in Seventy-Five :
Hardly a man is now alive
Who remembers that famous day and year.

He said to his friend, — " If the British march
By land or sea from the town to-night,
Hang a lantern aloft in the belfry-arch
Of the North-Church-tower, as a signal-light, —
One if by land, and two if by sea ;
And I on the opposite shore will be,
Ready to ride and spread the alarm
Through every Middlesex village and farm,
For the country-folk to be up and to arm."

Morse Code

A .--	J .-- --	S ...	1 .-- -- --
B -....	K -.-	T -	2 ..-- --
C -.-.-.	L .-...	U ..-	3 ...--
D -...	M --	V ...-	4 -
E .	N --.	W .--	5
F ..-..	O ---	X -...-	6 -.....
G --.-	P -....	Y -.-.-	7 ---....
H 	Q -.-.-	Z --...	8 ---....
I ..	R .-.	0 - - - - -	9 - - - - .

Properties of a good code

- Unambiguous
 - every distinct meaning gets a distinct string
 - possible to tell when a string ends (and the next begins)
- Noise-tolerant
 - e.g. if one bit is flipped, the intended message is still the 'closest' message
- Efficient
 - uses least number of symbols to communicate messages

Foundation:

Entropy

What makes a code efficient?

- Assumptions and notation:
 - Messages: n distinct message types, each message i occurs with probability p_i
 - Symbols: b distinct symbols in alphabet Σ , each equally costly to transmit
- Basic insight:
 - messages with higher probability should get shorter strings

Example

- Los Angeles weather
 - outcome 1: $p_1 = 335/365$
 - 23° C, mostly sunny, no earthquake
 - outcome 2: $p_2 = 20/365$
 - 23° C, mostly sunny, earthquake
 - outcome 3: $p_2 = 9/365$
 - 23° C, 15 minutes rain, no earthquake
 - outcome 2: $p_2 = 1/365$
 - 23° C, 15 minutes rain, earthquake

Cost of communicating the weather

- Code 1

- first position: 1 = 23 C
- second position: 1 = rain
- third position: 1 = quake

- Average cost over year:

- 3 bits/day
- (every day takes exactly 3 bits)

- Code 2

- [no message]: o_1
- 0: o_2
- 1: o_3
- 01: o_4

- Average cost over year:

- $(335*0 + 20*1 + 9*1 + 1*2)/365 = 0.08$ bits/day

N outcomes

- Suppose a binary alphabet
- Suppose n distinct, equiprobable outcomes
- E.g. suppose $n = 8$
 - at least one string must have length 3 (because $2^0 + 2^1 + 2^2 = 7$ distinct outcomes)
 - in fact, to reliably know when a message ends, the average length must be at least 3 (since $2^3 = 8$)
- The *average message length* must be at least $\log_2 n$
- If alphabet has b symbols, $\log_b n$

Formalizing the intuition

- Define *surprisal* of an event as follows:

$$I(o_i) = -\log_b p_i$$

- where b is the number of symbols in the alphabet
- Intuition: Efficient code will assign shorter codes to more frequent messages
- Formalization: If message has probability p_i , efficient code will assign a string of length $\lceil \log_b p_i \rceil$

Example: Fair coin

- Fair coin:
 - heads $p_1 = 1/2$
 - tails $p_2 = 1/2$
- Assume binary alphabet
 - $!(\text{heads}) = -\log_2 1/2 = -\log_2 2^{-1} = -(-1) = 1$
 - $!(\text{tails}) = \dots = 1$
- Each outcome has a surprisal of 1 bit

Entropy, or Uncertainty

- Formally, the *entropy* (or *uncertainty*) associated with a random variable is the expected surprisal.
- A discrete, idendepently-and-identically-distributed (i.i.d) RV can be defined as a triple $(X, \Omega, \text{Pr}_\mu)$ where Ω is the event space, Pr_μ is a probability distribution over Ω , and X is the current/next event

$$\begin{aligned} H[X] &= E[!(X)] = \sum_{\omega[i] \in \Omega} p_i \times !(\omega_i) \\ &= - \sum_{\omega[i] \in \Omega} p_i \log_b p_i \end{aligned}$$

Example

- Fair coin
 - heads $p_H = 1/2$ $-\log_2(p_H) = 1$
 - tails $p_T = 1/2$ $-\log_2(p_T) = 1$
- Uncertainty:
 - $H[X] = -\sum_{\omega \in \{H,T\}} p_\omega \log_2 p_\omega = -(1/2 \log_2 1/2 + 1/2 \log_2 1/2) = -(1/2 \times 1 + 1/2 \times 1) = -(-1/2 + -1/2) = -(-1) = 1$
- The uncertainty of a fair coin is one bit.
 - The average length of binary strings needed to communicate the outcome of a fair coin flip is 1.

Summary

- 'Information' is formalized as the *average number of symbols* needed to communicate the outcome of a discrete, i.i.d. random variable.
- Founded on the philosophical notion of 'reducing uncertainty about the state of the world'.
- The surprisal of an event is the negative log of its probability (with base b , often 2 for a binary code).
- Entropy/Uncertainty is the expected surprisal.

Implementation

```
# usage: python unigram_entropy.py [corpus_filename]

import sys, codecs, math
input_filename = sys.argv[1]

# read in word frequencies
def read_word_frequencies_to_dic(filename, enc='utf8'):
    word_frequency_dic = {}
    fin = codecs.open(input_filename, encoding=enc)
    for line in fin:
        for word in line.split():
            word_frequency_dic[word] = word_frequency_dic.get(word,0) + 1
    return(word_frequency_dic)

word_frequencies = read_word_frequencies_to_dic(input_filename)

# calculate entropy
def calc_entropy_from_frequencies(freq_dic):
    H = 0.0
    total_freq = float(sum(freq_dic.values()))
    for freq in freq_dic.values():
        prob = freq/total_freq
        H -= prob * math.log(prob, 2)
    return(H)

print calc_entropy_from_frequencies(word_frequencies)
```

robert@robert-Q550LF: ~/Desktop/PMCK

```
robert@robert-Q550LF:~/Desktop/PMCK$ head ~/Corpora/mobydick/mobydick.txt  
the project gutenber ebook of moby dick or the whale by herman melville
```

```
this ebook is for the use of anyone anywhere at no cost and with  
almost no restrictions whatsoever you may copy it give it away or  
re-use it under the terms of the project gutenber license included  
with this ebook or online at www.gutenberg.org
```

```
title moby dick or the whale
```

```
robert@robert-Q550LF:~/Desktop/PMCK$ python unigram_entropy.py ~/Corpora/mobydic  
k/mobydick.txt
```

```
10.0829571974
```

```
robert@robert-Q550LF:~/Desktop/PMCK$
```

Conclusion

- Entropy gives us a precise way to measure the amount of uncertainty in a process
- We do this by treating the process as a discrete, i.i.d. random variable.
- The uncertainty is the average length of a (binary) code needed to communicate the outcome of the RV on a trial.
- **Entropy provides a way to measure the information content of an RV.**

Extensions:

Conditional entropy, KL divergence, etc..

Practicality?

- Shannon's work on coding theory was of enormous practical relevance in designing electronic communications systems.
- It provided a principled way to estimate the information rate of a noisy channel, as well as a way to construct the most efficient code, guaranteeing that the theoretically maximum efficiency could almost be achieved.
- But how is this of use to **scientists**, rather than engineers?

Measuring relationships between two RVs

- The definition of entropy can be readily extended to multi-variate RVs.
- This is of enormous scientific value in *quantifying the informativity of one RV for another*.
- For example, one RV might represent an acoustic cue like VOT, and the other might represent a phonological parse.
- Or, one RV might represent the *theoretically predicted* distribution, and another might represent the *empirically observed* dist'n.

Entropy of two RVs

- Let (X, Ω_X, \Pr_X) and (Y, Ω_Y, \Pr_Y) be two RVs. Then the *joint entropy* is defined:

$$H[X, Y] = E[!(X, Y)]$$

$$= \sum_{(\omega[X], \omega[Y]) \in \Omega[X] \times \Omega[Y]} p_{(\omega[X], \omega[Y])} \times !(\omega[X], \omega[Y])$$

$$= - \sum_{(\omega[\xi], \omega[\psi]) \in \Omega[X] \times \Omega[Y]} p_{(\omega[X], \omega[Y])} \log_b p_{(\omega[X], \omega[Y])}$$

Example: Two fair coins

- Fair **red** coin (X):
 - $\Pr(\text{head}) = 1/2$
 - $\Pr(\text{tail}) = 1/2$
- Fair **blue** coin (Y):
 - $\Pr(\text{head}) = 1/2$
 - $\Pr(\text{tail}) = 1/2$
- (assume independence)

- Joint entropy

$$\begin{aligned} H[X, Y] &= -(p_{HH} \log_2 p_{HH} + p_{HT} \log_2 p_{HT} + p_{TH} \log_2 p_{TH} + p_{TT} \log_2 p_{TT}) \\ &= -4 \times (1/4 \log_2 1/4) = -4 \times 1/4 \log_2 2^{-2} = -(-2) = 2 \end{aligned}$$

Joint entropy

- In the previous example, $H[X, Y] = 2 = H[X] + H[Y]$
- More generally,
 - $H[X, Y] \leq H[X] + H[Y]$ *(triangle inequality)*
 - $H[X, Y] = H[X] + H[Y]$ if and only if X and Y are statistically independent
- The uncertainty in 2 fair coins is simply 2 times the uncertainty in a single fair coin

Joint entropy with non-independence

- However, when two RVs are not independent, the joint entropy is less than the sum
- Let X be the RV associated with flipping a coin, and let Y be the RV associated with the image of the same coin in a mirror.

	head	tail
head	1/2	0
tail	0	1/2

$$\begin{aligned} H[X, Y] &= -(p_{HH} \log_2 p_{HH} + p_{HT} \log_2 p_{HT} + p_{TH} \log_2 p_{TH} + p_{TT} \log_2 p_{TT}) \\ &= -(2 \cdot 1/2 \log_2 1/2 + 2 \cdot 0 \log_2 0) \\ &= -(-1) = 1 \end{aligned}$$

Joint entropy with non-independence

- In the preceding example, the outcome of one coin flip completely determines the outcome of the other.
- Thus, the 'real' amount of information in both coin flips is equal to 1 bit -- the uncertainty in a single coin flip.
- When two RVs are completely independent, the joint entropy is just their sum. When they are completely dependent, the joint entropy is equal to the entropy of just one. What about...

Conditional Entropy

- The most normal case is that two RVs are not independent, but also not co-determined. In this case, we may ask, how much information does one variable contribute to the other?
- The *conditional entropy* $H[Y | X]$ is the average amount of uncertainty remaining in Y when the value of X is known:

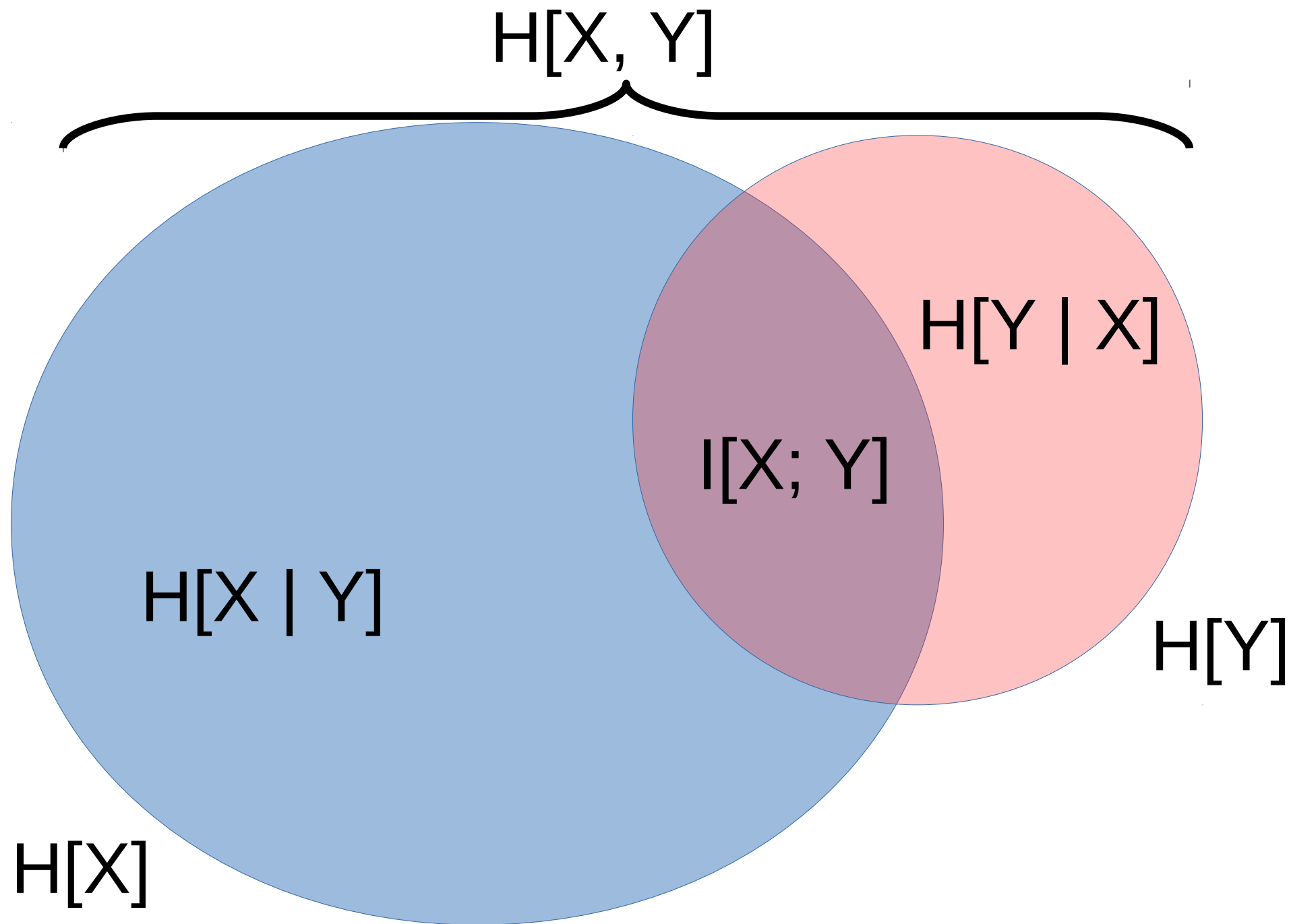
$$H[Y | X] = H[X, Y] - H[X]$$

Mutual Information

- The *mutual information* is a (symmetric) measure of the dependence between two RVs:

$$I(X; Y) = H[X] + H[Y] - H[X, Y]$$

A picture is worth... 10,082.9 bits



KL Divergence

- Often in science we have a theoretically predicted probability distribution p , and an empirically observed distribution q
 - (E.g. a logistic regression of some data)
- The KL divergence measures how many bits are wasted by using an optimal code for q when the messages are actually drawn from p

$$\text{KL}(p \parallel q) = \sum_{\omega[i] \in \Omega} p_i \times \log_b (p_i/q_i)$$

- Model-fitting is normally equivalent to minimizing KL divergence

Applications

Application 1

Morphological processing in
Serbian feminine nouns and
Dutch derivational morphology

Morphological processing

- Serbian feminine (Kostić, 1991, *et seq.*)
 - *trav-a* nom.sg + gen.pl
 - *trav-e* gen.sg + nom.pl + acc.pl
 - *trav-i* dat.sg + loc.sg
 - *trav-u* acc.sg
 - *trav-om* instr.sg
 - *trav-ama* dat.pl + instr.pl + loc.pl

Kostić: the problem

- The problem that a listener needs to solve is:
 - given a lexeme, identify the meaning
 - in Serbian, a lexeme consists of the root, followed by an inflectional exponent

Kostić: the solution

- For a stem L (e.g. *trav*), let
 - m index over inflectional exponents
 - so that L_m represents a lexeme (e.g. *trav-a*)
 - F_m = frequency of L_m
 - R_m = number of inflectional property sets for which L_m is the exponent (e.g. *trav-a*: 2 because this is both the nom.sg and gen.pl)
 - $p_m = F_m/R_m$
 - $P = \sum_m p_m$
 - $!(m) = -\log_2 (p_m/P)$

(Comments on Kostić)

- Confusing: the p_m values are almost equivalent to treating the paradigm as a joint RV, with “exponent (e.g. -a)” and “inflectional property set (e.g. nom.sg)” as the dimensions, and simply assuming equal probability for all IPS's
- Confusing: “accusative case in Serbian predominantly takes an object role but it can also denote time, place, purpose, cause, etc..”
 - so, does 'meaning' mean IPS, or IPS-semantics pair??

Kostić: results

- Kostić reports that the proposed surprisal measure accounts for over 90% of the within-paradigm variance in lexical decision reaction times (LDRTs)
- Moscoso del Prado Martin et al. (2004) report an enriched variant of Kostić's measure which handles both inflectional and derivational morphology
- The 'information residue' they propose outperforms a combination of simpler measures in predicting LDRTs for Dutch

Application 2

The evolution and organization
of morphological paradigms

The problem

- Ackerman, Mahloun, Blevins and colleagues are interested in the relation between enumerative complexity and the cell-filling problem
- *enumerative complexity* -- any measure of the richness of an inflectional system
- *cell-filling problem* -- how difficult is it to predict the exponent for an arbitrary cell in the paradigm, given the form of another arbitrary cell?

Examples

- Greek

	singular				plural			
<u>Class</u>	<u>nom</u>	<u>gen</u>	<u>acc</u>	<u>voc</u>	<u>nom</u>	<u>gen</u>	<u>acc</u>	<u>voc</u>
1	-os	-u	-on	-e	-i	-on	-us	-i
2	-s	-∅	-∅	-∅	-es	-on	-es	-es
3	-∅	-s	-∅	-∅	-es	-on	-es	-es
4	-∅	-s	-∅	-∅	-is	-on	-is	-is
5	-o	-u	-o	-o	-a	-on	-a	-a
6	-∅	-u	-∅	-o	-a	-on	-a	-a
7	-os	-us	-os	-os	-i	-on	-i	-i
8	-∅	-os	-∅	-o	-a	-on	-a	-a

Enumerative complexity

- Greek has 8 declension classes
- For each class, there are a variety of exponents
- For each inflectional property set, there are a variety of exponents
- During the course of language acquisition, a child must learn all of these to produce the language properly
- Ackerman & Mahloun discuss the extraordinary case of Chiquitlan Mazatec, which has 109 nominal declension classes!

The Cell-Filling Problem

- Informally: How hard is it to do the *wug* test?
- More formal: On average, how much uncertainty is there as to the form of an arbitrary cell i , given the form of a different arbitrary cell j ?
- (For example: you learned a new verb *google* from the 1st-person singular nonpast; now you want to use the same verb stem but in the 3rd-person plural past. How many options are there, and how confident are you in them?)

Ackerman & Mahlouf (2013)

- The *Low Entropy Conjecture*
 - “enumerative complexity is effectively unrestricted, as long as the average conditional entropy [of one cell given another] is low” (p. 436)
- *a priori* argument: cell-filling problem is what matters, not enumerative complexity
- typological study: enumerative complexity varies enormously across languages, but average conditional entropy of cells is never much more than 1 bit
- simulation: shuffled variant of Mazatec without morphological implicational relationships exhibits far higher average conditional entropy of cells

Application 3

The influence of orthography
on the adaptation of English
vowels in Korean loanwords

The adaptation of English vowels

- Problem: phonetic and phonological structures do not match
- Adaptation is usually to the closest phonetic match
- E.g. English has two high front vowels (c.f. *tin*, *teen*), both adapted as Korean high front vowel
- Stress has pervasive effects in English phonology, including vowel reduction
- Korean does not have an analogue of unstressed vowels; adaptation is quite variable

Orthographic influence?

<u>English</u>	<u>Source SR</u>	<u>Loan SR</u>	<u>Hangeul</u>
<u>a</u> cad <u>e</u> my	[ə ^h k ^h ædə ^h mi]	[a ^h k ^h adɛ ^h mi]	아카데미
<u>a</u> cro <u>p</u> olis	[ə ^h k ^h ɾap ^h əlɪs]	[a ^h k ^h ɪɾop ^h olɪsɪ]	아크로폴리스
<u>a</u> lu <u>m</u> in <u>u</u> m	[ə ^h luminə ^h m]	[a ^h l:uminj <u>u</u> m]	알루미늄
ba <u>l</u> le <u>r</u> ina <u>a</u>	[pæ ^h lɛ ^h ɾinə ^h]	[pa ^h l:ɛɾina ^h]	발레리나


How to measure the effect?

- Information theory to the rescue!
- Treat all the relevant vowels as an RV:
 - P: the phonological identity of the English vowel
 - e.g. [ə]
 - O: the orthographic identity of the English vowel
 - e.g. <a>
 - K: the adapted Korean vowel grapheme
 - e.g. 아

source O	a	ca	de	my
source P	[ə	k ^h æ	də	mi]
loan SR	[a	k ^h a	dɛ	mi]
loan K	아	카	데	미

	P	O	K
1.	ə	a	아
2.	æ	a	아
3.	ə	e	에
4.	i	y	이

source O a ca de my
source P [ə] k^hæ də mi]
loan SR [a k^ha dɛ mi]
loan K 아 카 데 미



	P	O	K
1.	ə	a	아
2.	æ	a	아
3.	ə	e	에
4.	i	y	이

source O	a	ca	de	my
source P	[ə]	k ^h æ	də	mi]
loan SR	[a	k ^h a	dɛ	mi]
loan K	아	카	데	미

	P	O	K
1.	ə	a	아
2.	æ	a	아
3.	ə	e	에
4.	i	y	이

Now (P, O, K) is a joint random variable.

We can use conditional entropy to measure the influence of orthography.

Logic

- $H[K | P]$
 - the amount of uncertainty that remains in the Korean vowel adaptation, given knowledge of the phonetic identity of the English source vowel
- $H[K | P, O]$
 - the amount of uncertainty that remains in the Korean vowel adaptation, given knowledge of both the English spelling *and* vowel identity
- orthographic information gain: $H[K | P] - H[K | P, O]$
 - cannot just measure $H[K | O]$ because P, O correlated

Results

<u>stress</u>	<u>H[K P]</u>	<u>H[K P,O]</u>	<u>O-gain</u>	<u>chance</u>
primary	1.08	0.69	0.39	0.35±.03
none	1.71	0.70	1.01	0.30±.04

(NB *Chance* was measured by a scrambling operation, in which the associations between **O** and **P**, **K** were scrambled, guaranteeing that $H[K | P,O]$ cannot yield 'true' improvement over $H[K | P]$. Owing to finite sampling effects, the statistic still comes out positive -- we take that as the baseline/chance level.)

Interpretation

- The *orthographic information gain* is defined as the extra information that English spelling must contribute to Korean vowel adaptation, beyond the English vowel identity.
- Daland & Oh (under revision) show that this quantity is positive for all stress levels.
 - but it is greatest in both absolute and relative magnitude for unstressed vowels
 - the corresponding value of *phonological information gain* shows the opposite pattern
- Phonetics matters more for stressed vowel adaptation, orthography matters more for unstressed adaptation

Summary and Conclusions

What is Information Theory?

- A statistically rigorous formalization of the intuition:
- *Information is anything that reduces our uncertainty about the state of the world. The amount of information is measured by the reduction in uncertainty.*
- Anytime a phenomenon can be characterized by a (set of) random variable(s), information theory is a potential means for analysis.

What is information good for?

- In scientific applications, information theory is most useful for characterizing the relationship *between* two or more RVs.
 - morphological processing: conditional entropy of an inflectional exponent predicts lexical decision RT
 - evolutionary morphology: average conditional entropy of one paradigm cell given another ('cell-filling problem') constrains morphological system, not absolute entropy of the paradigm ('enumerative complexity')
 - loanword adaptation: show that the association between English vowel orthography and Korean vowel adaptation is higher than predicted by chance/vowel identity alone

Thank you!