Natural Languages

1. Structural information is more important when considering processing difficulty. Long-distance dependencies within a construction are significantly correlated to how long a human will take to parse a sentence. Whereas Zipf's law suggests that infrequent constructions/words are commonplace in natural language, and so are not on their own a reliable indicator of processing difficulty. Furthermore, structurally difficult-to-parse constructions are likely to be very infrequent in a corpus, but the inverse relationship does not hold (i.e., not all infrequent constructions are structurally difficult). Therefore, the structural information alone likely contains much of the information to be found from the frequency information.

2. Examples:

   I said (that) I didn't (did not) steal the jewellery.
   All (of) the time I'm (I am) thinking of ways *that I can* steal the jewellery.

   (In these examples, the bracketed words are omitted because they exist in low-information-frequency parts of the sentences, whereas the starred words are more verbose than necessary because they exist in high-information-frequency parts of the sentence)

   Counter-Examples:

   The fox *that was* crawling past my bed told us *that* we were special.

   (In this example, the starred words sound natural to include even though they create significant drops in information rate)

Formal Models and Learnability

(See attached)

Information Theory

1. (See attached)

2.

```
#!/usr/bin/python3.9

import numpy as np
import warnings

with open("alice.txt") as f:
    data = f.read()

alphabet = list(set(data))
```

```
alphabet_idx_lookup = {}
for i, char in enumerate(alphabet):
    alphabet_idx_lookup[char] = i


# Generate bigram probability matrix
bigram_matrix = np.zeros((len(alphabet), len(alphabet)), dtype=float)
first_letter_matrix = np.zeros(len(alphabet), dtype=float)
all_letter_matrix = np.zeros(len(alphabet), dtype=float)

last_char = None
for char in data:
    char = alphabet_idx_lookup[char]

    if last_char is not None:
        bigram_matrix[last_char, char] += 1

    all_letter_matrix[char] += 1
    if last_char is None or alphabet[last_char] == " ":
        first_letter_matrix[char] += 1

    last_char = char

# Normalize distributions
bigram_matrix /= np.sum(bigram_matrix, axis=1)[:, np.newaxis]
first_letter_matrix /= np.sum(first_letter_matrix)
all_letter_matrix /= np.sum(all_letter_matrix)

# Helpful pre-computations
with warnings.catch_warnings():
    # Catch divide-by-zero warnings
    # If the value really is zero, we'll never pick it anyway
    warnings.simplefilter("ignore")
    H_cond = -np.log2(bigram_matrix) * bigram_matrix
    H_first = -np.log2(first_letter_matrix) * first_letter_matrix
    H = -np.log2(all_letter_matrix) * all_letter_matrix

def generate_word():
    # Pick the first letter of the word according to first_letter_matrix
    current = np.random.choice(np.arange(len(alphabet)), p=first_letter_matrix)
    word = ""

    entropy = H_first[current]

    while (alphabet[current] != " "):
        word += alphabet[current]
        # Pick the next letter according to bigram_matrix
        nxt = np.random.choice(np.arange(len(alphabet)), p=bigram_matrix[current, :])

        # Add to entropy in accordance with the chain rule
        # NOTE: Assuming P(X_n|X_n-1, X_n-2, ..., X_1) = P(X_n|X_n-1)
        #       (Markov assumption)
        entropy += H_cond[current, nxt]

        current = nxt


    return word, entropy

words, entropies = zip(*list(generate_word() for _ in range(10000)))

# Sort words by entropy
words = list(np.array(words, dtype="object")[np.argsort(entropies)[::-1]])

print(words[:10])

# ['lokendinfisemakitheerminthaing', 'soprerokishaimucerulitinlishormad', 'ssatheastomaishersemeridouryo',
'thtoundshengrathechithey', 'appouteyithithemigucryphale', 'ishooushewilinioupotoumits',
'waingoulickithicoutim', 'waderplousoulingereeme', 'hesklloflriteryoventhntllld', 'sentheesearithailyore']
```

3.

a. The answer to the question can be considered the input, and the question itself can be considered the output after passing the answer through a noisy channel. Therefore, the answer with maximum likelihood is the one which maximises P(answer)P(question|answer)

b. The input to the channel can be some encoding of the "sense" of the word, and the output will be the word in context. The correct sense of the word will be the one which maximises P(sense)P(word in context|sense).

Y2021P7Q5

a-e. (See attached)

f. One hypothesis might be that the amount of time it takes a human to process a construction of the form ab^kc^jd (e.g., a highly highly highly contagious contagious virus) does not significantly depend on k or j above some threshold, perhaps 2. I believe this to be the case because past that threshold, the human would stop reading each word individually, but instead recognise it as adding no meaning, only emphasis, and can quickly find where the repetition ends.

Another hypothesis might be that humans might process language more like the inductive method from part (c) than the grammatical method in part (a). I believe that humans are more likely to think that, for example. "contagious" can be replaced with "highly contagious" in a valid sentence and the sentence will still be valid, as opposed to thinking that the word "contagious" needs specific types of words before and after it.

Formal Models & Learnability

1. By considering the incomplete inductions for "abc"

$$\dfrac{\dfrac{a}{x}R \qquad \dfrac{\dfrac{b}{x}R \quad \dfrac{c}{?}R}{s\backslash x}<}{s}<$$

it is clear that $? \to s\backslash x\backslash x$

$\therefore (c, s\backslash x\backslash x) \in R$

Likewise for "abdc"

$$\dfrac{\dfrac{a}{x}R \qquad \dfrac{\dfrac{b}{x}R \quad \dfrac{\dfrac{d}{?}R \quad \dfrac{c}{s\backslash x\backslash x}R}{s\backslash x\backslash x}<}{s\backslash x}<}{s}<$$

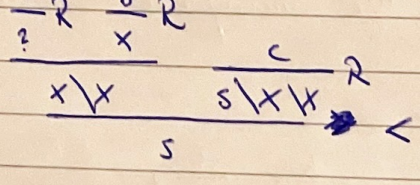
$\therefore ? \to s\backslash x\backslash x / (s\backslash x\backslash x)$

$\therefore (d, s\backslash x\backslash x / (s\backslash x\backslash x)) \in R$

Likewise for "ebc"

$$\dfrac{\dfrac{e}{?}R \qquad \dfrac{\dfrac{b}{x}R \quad \dfrac{c}{s\backslash x\backslash x}R}{s\backslash x}<}{s}$$

or



$\therefore ? \rightarrow X$ or $X \backslash X \backslash X$

$\therefore$ either $(e, x) \in R$
or $(e, x \backslash x \backslash x) \in R$

2. Let x_i = the i^{th} string in the language, for some ordering.

$x_i = (x_i^1, x_i^2, x_i^3, \dots, x_i^{n_i})$ where $x_i^j \in \Sigma$ and n_i is the number of symbols in x_i

Let $(x_i^j, X_i^j) \in R$ for all $1 < j \leq n_i$

$(x_i^1, S/X_i^2/X_i^3/\dots/X_i^{n_i}) \in R$

If there are N strings in the language, the maximum number of types a symbol could be given is $\sum_{1 \leq i \leq N} n_i$ (e.g., if every string consists of the same symbol over and over again)

Let $k = \sum_{1 \leq i \leq N} n_i$.

Note: n_i is finite and N is finite, $\therefore$ k is finite

The language is a k-valued C_{cg}, $\therefore$ it is learnable within Gold's paradigm.

Information Theory

1.

x_i	y_i	$P(x_i)$	$P(y_i)$	$P(x_i \wedge y_i)$	$P(x_i \mid y_i)$
0	0	q	$q(1-p)+p(1-q)$	$q(1-p)$	$1-p$
0	1	q	$qp+(1-p)(1-q)$	qp	p
1	0	$1-q$	$q(1-p)+p(1-q)$	$(1-q)p$	p
1	1	$1-q$	$qp+(1-p)(1-q)$	$(1-q)(1-p)$	$1-p$

Let $X = (x_1, x_2, \ldots, x_n)$

$Y = (y_1, y_2, \ldots y_n)$

Assume large n.

$$H(X) = -\sum_{i=1}^{\hat{}} P(x_i) \log_2 P(x_i) \approx n P(x=0)^2 \log_2 P(x=0) + n P(x=1)^2 \log_2 P(x=1)$$
$$= nq^2 \log_2 q + n(1-q)^2 \log_2 (1-q)$$

$$H(X|Y) = -\sum_{i=1}^{n} P(x_i|y_i) \log_2 P(x_i|y_i)$$

$$\approx -n P(x=0 \wedge y=0) P(x=0|y=0) \log_2 P(x=0|y=0)$$
$$- n P(x=0 \wedge y=1) P(x=0|y=1) \log_2 P(x=0|y=1)$$
$$- n P(x=1 \wedge y=0) P(x=1|y=0) \log_2 P(x=1|y=0)$$
$$- n P(x=1 \wedge y=1) P(x=1|y=1) \log_2 P(x=1|y=1)$$

$$= -n \left(q(1-p)^2 \log_2 (1-p) + qp^2 \log_2 p + (1-q)p^2 \log_2 p + (1-q)(1-p)^2 \log_2 (1-p) \right)$$
$$= -n \left((1-p)^2 \log_2 (1-p) + p^2 \log_2 p \right)$$

$$I(X;Y) = H(X) - H(Y)$$
$$\approx -n \left(q^2 \log_2 q + (1-q)^2 \log_2 (1-q) - p^2 \log_2 p - (1-p)^2 \log_2 (1-p) \right)$$

$$\therefore \frac{d}{dq} I(X;Y) \triangleq -n\left(2q\log_2 q + \frac{q}{\ln 2} - 2(1-q)\log_2(1-q) - \frac{1-q}{\ln 2}\right)$$

$$= -n\left(2q\log_2 q - 2(1-q)\log_2(1-q) + \frac{2q-1}{\ln 2}\right)$$

~~$\therefore \log_2\left(q^{2q}\right) = \log_2\left((1-q)^{2(1-q)}\right) \cdot e^{\frac{1-2q}{\ln 2}}$~~

~~$\therefore q^{2q} = (1-q)^{2(1-q)} \cdot e^{1-2q}$~~

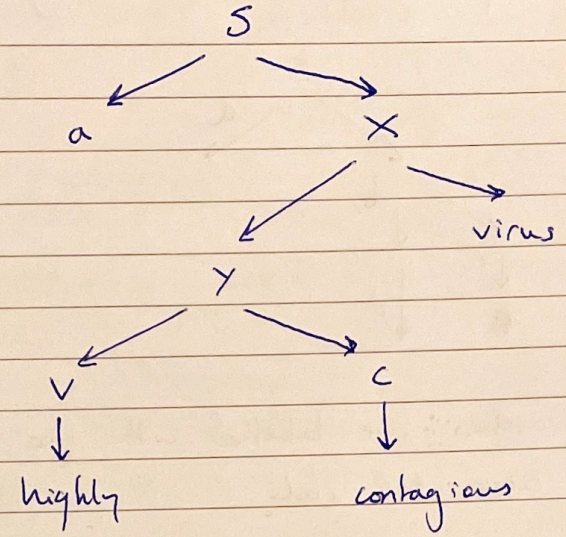
Observe:

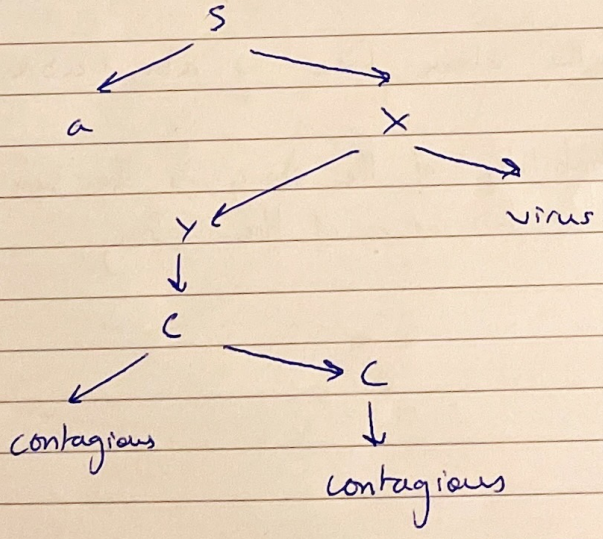
$$\frac{\partial}{\partial q}\Big|_{q=1/2} I(X;Y) \triangleq -n(-1+1+0) = 0$$

$\therefore q = 1/2$ is a maximum of $I(X;Y)$

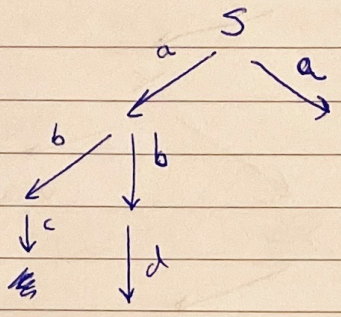
Y2021P705

a.


S
a
X
virus
Y
V
C
highly
contagious


S
a
X
virus
Y
C
C
contagious
contagious

b. Consider a parse tree of the string as below



where branches are labelled with the probabilities of the associated rule.

The probability of the derivation is calculated by multiplying the probabilities of each rule along each path from root to leaf, and summing the products.

Eg, in the above tree, $p: abc + abd + a$

The probability of the string is the sum of the probabilities of each derivation of the string.

c. $L_1 = \emptyset$

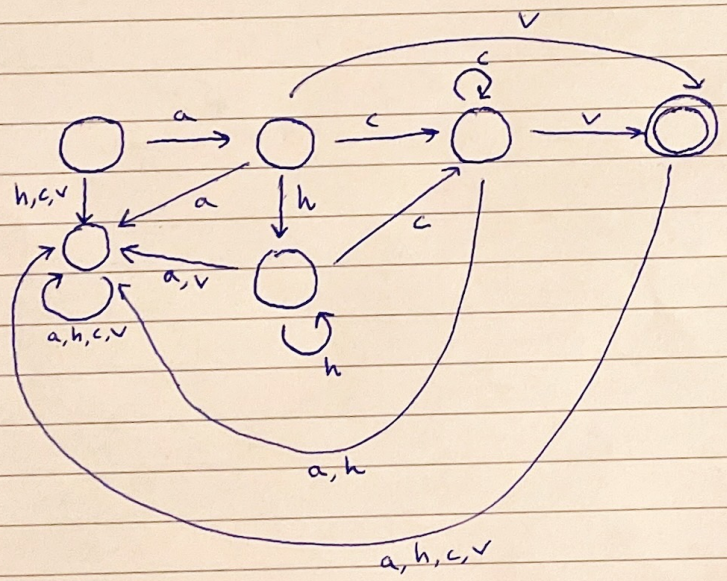
$L_2 = \{a, v\}$

$L_3 = L_2 \cup \{acv\}$

$L_4 = L_3 \cup \{ahcv, accv\}$

$= \{av, acv, ahcv, accv\}$

d. $a((h^*c)|\epsilon)c^*v$



e. ~~$H(X) = -\sum_{x \in \Sigma} \sum_{y \in \Sigma} p(x,y) \log_2 p(x|y)$~~

(I am confused by ~~this~~ this question — I'm not sure what X is supposed to be conditional on)