Formal Languages Example Sheet 2

OTOC: These exercises are bookwork. Do these only if you want to check your answers (Only TO Check).
OIFT: These exercises provide deeper insights or more practice (and are exam-level questions), but if you don't have time do not attempt them. (Only IF Time).
Optional: Attempt these only if you want more practice.
Advanced Optional: Attempt these only if you want a challenge or after the exams are over.

Information theory

Core exercises

Exercise 1 [2-symbol entropy] Write the formula for the entropy of a two-symbol $\{0,1\}$ random variable X with $\mathbb{P}(X=1)=p\in[0,1]$.

(a) Plot the entropy as a function of p.

(b) What happens at $p=0$ or $p=1$?

(c) Find the points where entropy is minimised/maximised.

(d) (**Optional**) For which discrete distribution over n symbols is the entropy maximised? (*[**Hint:** Use Lagrange multipliers]*)

[Example Sheet 2 Information Theory 1]

Exercise 2 [Chain rule]

(a) Prove that $H(Y\mid X)+H(X)=H(X,Y)$.

(b) Compute $H(X^2\mid X)$.

Exercise 3 [Kullback-Leibler divergence]

(a) Prove that $I(X;Y)=D(p(x,y)\parallel p(x)p(y))$.

(b) (**OIFT**) Demonstrate that there exist distributions p and q, such that $D(p\parallel q)\neq D(q\parallel p)$. Why is this a disadvantage?

(c) (**Optional**) Can you think of a way to fix this problem?

Exercise 4 [Coding] Consider symbols A,B,C,D,E with probabilities:

$$p_A=0.3, p_B=0.3, p_C=0.2, p_D=0.1, p_E=0.1$$

(a) What is the property that a coding scheme must have to ensure unique decodability for multiple symbols?

(b) Design a coding scheme that minimises the expected length of the code.

(c) What is the expected length of your code and how does it compare to the entropy?

Exercise 5 [Entropy rate]

(a) Show that for independent random variables X_1 and X_2, $H(X_1,X_2)=H(X_1)+H(X_2)$ and give an interpretation in terms of mutual entropy.

(b) Define entropy rate for a collection of random variables.

(c) Show that for i.i.d. random variables X_i, the entropy rate is equal to $H(X_1)$.

(d) What is the entropy rate of the sentences produced by a typewriter that types uniformly at random?

Exercise 6 [Surprisal]

(a) (**OTOC**) Define *surprisal*.

(b) How might it be used in text analysis?

(c) (**Optional**) Experiment with surprisal in `mains/ngram_demo/ngram_main.cc`. Try running `alice_word_unigram()` and `alice_char_unigram()`. Remember to set `alice_file_location` at the top of the file (or pass it as a flag).

Exercise 7 [Language modelling] (**OIFT**)

(a) Define the *unigram language model*.

(b) What are the problems with the unigram character and word language models?

(c) Define the *bigram language model*.

(d) How does this model compare with the bigram language model? Are there any issues with this model?

(e) (**Optional**) Compare the unigram and bigram language models with any of the following modern language models: i) demo, ii) demo, iii) demo. You can try assessing their geography knowledge (e.g. capital cities), programming language fluency, story-telling, linguistic knowledge (garden path sentences) etc.

(f) (Open-ended) Is frequency information or structural information more important when considering processing difficulty?

[Example Sheet 2 Natural Languages 1]

Exercise 8 [Noisy channel]

(a) (**OTOC**) Define the *noisy channel framework*.

(b) What is the connection between the noisy channel framework and Bayes' rule?

(c) Explain how you could frame the following tasks as noisy channel problems:

i. Disambiguating multiple senses of a word.

ii. (**OIFT**) Automatically answering questions.

iii. (**OIFT**) Spelling correction.

iv. (**OIFT**) Machine translation.

[Example Sheet 2 Information Theory 3]

Exercise 9 [Binary symmetric channel] A *binary symmetric channel* is one where the input x_i and the output y_i are both in $\{0, 1\}$. The channel is characterised by p, the probability that an input bit is transmitted as the opposite bit. If q is the probability that the source sends $x = 0$ and $1 - q$ the probability of $x = 1$, show that the mutual information is maximised when zeros and ones are transmitted with equal probability (i.e. when $q = 0.5$).

Exercise 10 [Practical] Using the processed Alice in Wonderland file, write some simple code to generate some good candidates for nonsense words by:

(a) Finding the probability distribution defined by a bigram language model. ***[Hint:*** *Count occurrences of unigrams and bigrams.]*

(b) Generating some words using the probability distribution.

(c) Selecting the 10 words whose information rate is lowest.

[Hint: *In case you need help, you can look at the code in (and referenced by)* `mains/ngram_demo/ngram_main.cc`*.]*

[Example Sheet 2 Information Theory 2]

Paper-based exercises

These exercises are based on the papers whose results were presented in the lectures. You do not need to read the papers to answer these questions. Sometimes reading the introduction (or the abstract) can give you some insights.

Exercise 11 ["Word lengths are optimised for efficient communication"] You can find the paper here.

(a) How is the context of a word represented in this study?

(b) How is the information conveyed by a word measured?

(c) What steps would you follow to produce the figure on **Lecture 6 slide 12**?

(d) What steps would you follow to produce the figure on **Lecture 6 slide 13**?

(e) How do these figures support the main claim of the paper?

Exercise 12 ["Entropy rate constancy in text"] You can read this presentation for more details (or the original papers, but these are much longer).

(a) What is the *entropy rate constancy* assumption? How does it affect utterances? Give examples.

(b) (Open-ended) Do you agree with this principle? What factors affect the channel capacity?

Exercise 13 ["The communicative function of ambiguity in language"]

(a) According to this theory why is *language ambiguous*? Give your own examples.

(b) Do you agree with the premise? Can you think of other reasons for ambiguity in language?

(c) What steps would you follow to produce the figure on **Lecture 7 slide 7**?

Exercise 14 ["A noisy-channel account of crosslinguistic word-order variation"]

(a) According to the paper, why does SVO have a better chance of preserving information? What evidence do they provide? Do you agree?

(b) According to the paper, why do people use SOV for inanimate objects?

Language learnability

Exercise 15 Consider a rigid classic categorial grammar $C_{cg} = (\Sigma, Pr, S, R)$, where $Pr = \{S, X\}$ and $S = S$. If a, b have type X and you know that $abc \in \mathcal{L}(C_{cg})$, $abdc \in \mathcal{L}(C_{cg})$, $ebc \in \mathcal{L}(C_{cg})$, give possible types for each of c, d and e.

[Example Sheet 2 Formal Languages and Learnability 1]

Exercise 16 Design a classic categorial grammar that can generate the following sentences:

- *She saw the elephant with the telescope.*
- *She drove down the street in the car.*
- *The man and the woman walked.*

Exercise 17 [Equivalence of CGs and CFGs] (**OIFT**) Show that context-free grammars and classic categorial grammars are equivalent.

Project 1 [Equivalence of CCG and TAGs] (+) Read this paper to see why CCGs are equivalent to TAGs.

Exercise 18 Explain why a finite class of finite languages is learnable within Gold's paradigm.

[Example Sheet 2 Natural Languages 2]

Exercise 19 Describe a learning paradigm where a learner could learn from two sources simultaneously (a bilingual).

[Example Sheet 2 Natural Languages 3]

Distributional models

Core exercises

Exercise 20 [Distributional representations]

(a) What is a *distributional representation*?

(b) The embedding projector is a visualisation tool for word embeddings. Search for the nearest neighbours of some example words. Do you empirically verify that similar words have word vectors that are close? What about antonyms?

(c) Attempt **[2014P9Q8 (a),(b)]**.

(d) (**Optional**) **[2018P9Q10 (a),(d)]**.

Exercise 21 [CBOW/Skipgram]

(a) (**Optional**) Explain how the *CBOW model* works.

(b) Explain how the *skipgram model* works.

(c) (Open-ended) How would you initialise the word embeddings?

(d) (Open-ended) How many dimensions would you use?

Exercise 22 Describe how you might use word distributions to compare the similarity of two characters in a text. What might any *similarity* be telling us about the characters?

[Example Sheet 2 Distributional Models 1]

Practical exercise

Project 2 You can go through the PyTorch tutorial on word embeddings (or some other tutorial) to see how wor2vec algorithms are implemented.

Additional optional exercises

Project 3 [Non-negative of KL] Jensen's inequality states that for a convex function f and X being a random variable,

$$E[f(X)] \geq f(E[X])$$

(a) By an appropriate choice of f show that $D(p \parallel q) \geq 0$ for any distributions p and q.

(b) Show that mutual information is also non-negative. Deduce that $H(X) \geq H(X \mid Y)$ for any random variables X and Y.

Project 4 [Huffman encoding] Read about Huffman encoding and its related algorithm. How would you apply it to the coding exercise above?