Informatics 1 Cognitive Science – Tutorial 3

Frank Keller, Carina Silberer, Frank Mollica

Week 4

1 Word Segmentation

Last week, we discussed aspects of language development in class, specifically speech segmentation and Bayes Rule. The goal of this tutorial is to revise what you have learnt by performing practical exercises.

1.1 Statistical Regularities

In class, we talked about transitional probability as a means to find word boundaries. Transitional probability is the *conditional probability* of adjacent elements. Conditional probability is defined as:

$$P(y|x) = \frac{p(x,y)}{p(x)} \tag{1}$$

and measures the probability of an event y under the assumption that another event x has happened. For example, y might correspond to the word *are* and x to the word *we*, so $P(y|x)$ would be the probability of *are* following *we*. The term $p(x,y)$ is the *joint probability* of x and y – it measures the probability of the occurrence of both events, x and y. As you learnt in the lecture, transitional probability is estimated as:

$$P(y|x) = \frac{p(x,y)}{p(x)} \approx \frac{freq(x,y)}{freq(x)}. \tag{2}$$

Exercise 1 You are given the sequence: $\boxed{\text{thenimmasawthenimbleanimal}}$

Table 1 contains the transitional probabilities computed for each letter bigram on the basis of the frequencies given in Table 2. For example, the first entry of Table 1 (.14) is the probability that a space (′ ′) will be followed by t, i.e., $P(t|' ')$. The second entry gives the probability that t will be followed by h, i.e., $P(h|t) = .32$, and so.

Table 2 should be read as follows: each entry corresponds to the number of times two adjacent letters occur in an underlying text. For example, the cell coloured in grey gives the occurrence frequency of the sequence *am*, i.e., $freq(a,m) = 245$. The last column titled *total* gives the frequencies of single letters (unigrams) as counted in the text. For example, a occurred 9615 times.

Determine the segmentation of the given sequence using transitional probabilities as cues. Do this by filling in the missing values in Table 1 by means of the frequencies given in Table 2. Then complete the chart in Figure 1 and insert the word boundaries.

''	t	h	e	n	i	m	m	a	s	a	w	t	h	e	n	i	m	b	l	e	a	n	i	m	a	l
–	.14	.32		.09		.04		.15	.11	.04	.02		.32		.09	.03	.04		.16	.16	.05		.03	.04	.15	.07

Table 1: Transitional probabilities between each pair of letters.

	''	t	h	e	n	i	m	a	s	w	b	l	total
''	0	4123	1879	578	597	2039	1416	3176	1955	1918	1150	836	28726
t	2591	286	3685	1111	11	674	66	340	164	60	0	134	11394
h	676	269	0	3106	5	1025	5	1296	16	0	6	10	7241
e	4807	407	17	458	1341	111	293	687	857	106	34	468	15251
n	1806	691	8	708	68	231	4	188	313	6	97	81	8438
i	632	1206	0	320	1983	2	307	67	1002	0	90	365	8278
m	357	1	0	764	17	254	38	465	82	0	59	5	3196
a	702	1290	7	2	2089	442	245	0	1070	188	197	625	9615
s	2425	945	288	943	16	451	51	309	355	37	2	58	7482
w	245	2	440	354	118	515	0	682	42	6	4	17	2886
b	12	17	1	607	3	76	1	91	26	1	1	197	1801
l	596	77	0	780	5	543	13	507	46	9	3	725	4843

Table 2: Letter bigram frequencies (source: The strange case of Dr Jekyll and Mr Hyde).

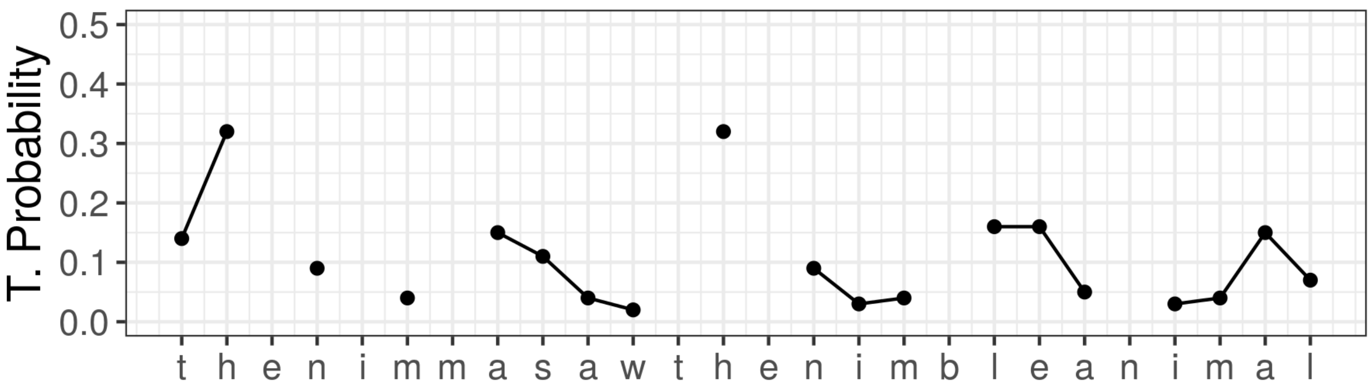


Figure 1: Transitional probabilities for the sequence *thenimmasawthenimbleanimal*.

1.2 Minimum Description Length

Exercise 2 In the lectures you also discussed the Minimum Description Length (MDL). Remember in MDL, the size(description) = size(lexicon) + size(data-encoding), where the size is equal to the number of characters (letter and digits, excluding spaces). Below are given three input sequences and two possible segmentations corresponding to each input.

1. Which segmentation hypothesis do you think will be favoured by the MDL model?

2. Compute the MDL for the segmentation hypotheses. Which hypothesis is favoured by the MDL model?

3. The two given segmentations of *thenimmasawthenimbleanimal* are both incorrect, (the correct one is *then imma saw the nimble animal*). Furthermore, the correct segmentation is one of many possible segmentations, for two of which you computed the MDL. What needs to be done to find the correct segmentation, assuming it will be the one with the least MDL?

4. What do you think is a better cue for word segmentation – transitional probabilities or MDL?

Input	Segmentation 1	Segmentation 2
thenimmasawthenimbleanimal	the nim ma saw the nim ble a nim al	the nimma saw the nimble animal
thenimmasaw the animal	the nim ma saw the a nim al	the nimma saw the animal
saw the cuteanimal	saw the cute a nim al	saw the cute animal

2 Bayesian Modeling

Last week, we discussed Bayesian Modeling as a way of capturing human reasoning and decision making. In this exercise, we will look at an example of how we can formalize a (simple) cognitive process in Bayesian terms.

Exercise 3 In an experiment on face recognition, subjects are presented with images of people they know, and asked to identify them. The images are presented for a very short period of time so that subjects may not have time to see the details of the entire face, but are likely to get a general impression of things like hair color and style, overall shape, skin color, etc. In this question we will consider how to formulate the face recognition problem as a probabilistic inference model.

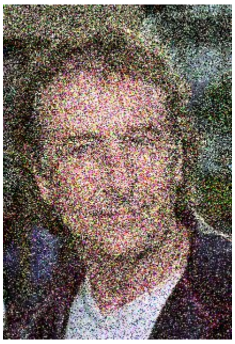
1. What is the hypothesis space in this problem? Is it continuous or discrete? Finite or infinite?

2. What constitutes the observed data d and what kinds of values can it take on?

3. Write down an equation that expresses the inference problem that the subjects must solve to identify each face. Describe what each term in the equation represents.

4. What factors might influence the prior in this situation?

5. Suppose one group of subjects sees clear images, such as the one on the left below, and another group sees noisy images, such as the one on the right below. Which term(s) in your equation will be different for the noisy group compared to the clear group?

6. What does the model predict about subjects' performance with noisy images compared to clear images?