Informatics 1 Cognitive Science – Tutorial 3 Solutions

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 and Solution You are given the sequence: 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., $P(t|\text{' '})$. The second entry gives the probability that *t* will be followed by *h*, i.e., $P(h|t) = .32$, and so on.

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	.43	.09	.03	.04	.01	.15	.11	.04	.02	.0007	.32	.43	.09	.03	.04	.02	.16	.16	.05	.22	.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).

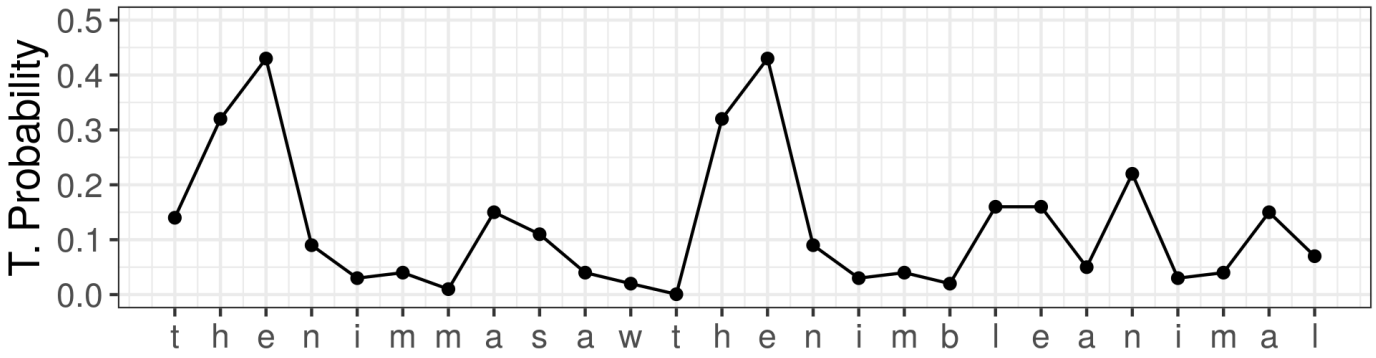


Figure 1: Transitional probabilities for the sequence *then|im|masaw|then|im|ble|an|imal*.

1.2 Minimum Description Length

Exercise 2 and Solution In the lectures you also discussed the Minimum Description Length (MDL). 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? ⇒ MDL minimises the length of words, as shorter words are more plausible → favours segmentation 1. But MDL minimises the number of different words (types) and maximises the probability of each word (prefers fewer words) → favours segmentation 2.

2. Compute the MDL for the two segmentation hypotheses. Which hypothesis is favoured by the MDL model?
   ⇒ See below. Segmentation 2 is favoured as its length is smaller than the length of segmentation 1.

3. The two given segmentations of *thenimmasawthenimbleanimal* are both not the correct one, which would be *then imma saw the nimble animal*. Furthermore, the correct segmentation is one of many possible segmentations, for two of which you computed their length. What needs to be done to find the correct segmentation, assuming it will be the one with the least length?

⇒ The search space comprising alternative possible segmentations needs to be explored. That involves systematically inserting word boundaries at different positions and measuring the length of the resulting segmentation. See Brent & Cartwright (1996, pp. 106–108), for a possible search algorithm.

4. What do you think is a better cue for word segmentation – transitional probabilities (TP) or MDL?
   ⇒ Open for discussion. Both approaches are not mutually incompatible. We could use transitional probabilities to suggest possible segmentations, and then use MDL to choose the best one.

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

Lexicon 1	Lexicon 2
1 the 2 nim 3 ma 4 saw 5 ble 6 a 7 al 8 cute	1 the 2 nimma 3 saw 4 nimble 5 animal 6 cute

Derivation 1	Derivation 2
1 2 3 4 1 2 5 6 2 7	1 2 3 1 4 5
1 2 3 4 5 6 2 7	1 2 3 1 5
4 1 8 6 2 7	3 1 6 5

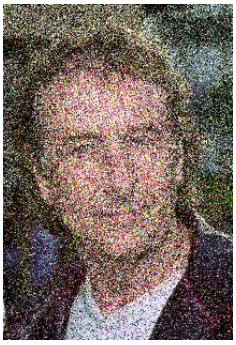
Length 1: 29+24 = 53	Length 2: 33+15=48

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?

Solution for Exercise 3

1. The hypothesis space is the set of different people that the subject knows, a finite (though very large) discrete space.

2. The data is the image the subject sees, or more precisely, what the subject actually perceives. It's difficult to say precisely what that might be without knowing more about the low-level features used by the visual system, but it could be things like the color and intensity in different regions of the image. In this case the values of the observed data would be continuous. However, if we were actually going to model this problem, we might want to simplify by assuming that higher-level discrete features are directly observed, e.g. face shape, hair style, skin color. But note that not all of these higher-level features would necessarily be observed for each trial. (We could also make an intermediate assumption, using a discretized space of color/intensity features, or maybe intermediate-level features like edges).

3. $$P(H|d) = \frac{P(d|H)P(H)}{P(d)} \tag{3}$$

   where the $P(H)$ is the subject's prior belief that any particular person will appear in the photo, $P(H|d)$ is the subject's posterior belief that any particular person is in the image, given what the subject sees in the image, $P(d|H)$ (likelihood) is the probability that a particular set of features will be perceived given that a particular person is shown in the image, and $P(d)$ is the overall probability of perceiving a particular set of features.

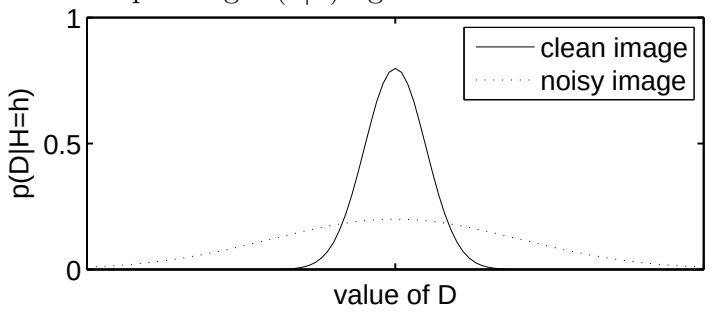
4. The prior could be influenced by factors such as the frequency or recency with which the subject has seen each of the people outside of the experimental situation, and the frequency of the particular person's face within the experimental situation (if images are reused). One could imagine other possible factors such as the emotional closeness of the subject to the person, or the frequency with which the subject has seen photographs of the person (as opposed to the live person).

5. The prior will be the same. The likelihood will be different, since the manipulation changes how the images look – there will be a different distribution over observed features given the same person being shown. $P(d)$ will be different since there is a different overall distribution of what the images look like. And since $P(d)$ and $P(d|H)$ are different, the posterior will also be different.

   In preparation for the next question it helps if we are more specific about the changes we expect in the likelihood. Consider the distribution $P(d|h)$ for a specific hypothesis h (say, Eric Idle). If the images are clear, then we would expect relatively little variation in the features we perceive when shown his image. That is, a relatively small number of possible values for y will have high

probability, and other possible values will have low probability. However, if the images are noisy, this effectively spreads out the probability mass over a larger number of possible values for y: we are more likely to see features that are further from those in the original image, but less likely to see features that are exactly those in the original image.

The above description assumes discrete values for d, but we can also get an intuition for what's going on by imagining that the different possible values for d are continuous values along a 1-dimensional space, and then plotting $P(d|h)$ against d:



The mean of these curves represents the "average" data that would be seen when Eric's picture is shown. The curves for the noisy and clean cases have the same mean, but the distribution of observations in the noisy case is broader than that in the clean case. [I'm assuming that the noise itself is unbiased, otherwise the mean could change also, but this would needlessly complicate the analysis.]

6. The model predicts that subjects will have a harder time discriminating Idle from Cleese in the noisy scenario. This is because the likelihood distribution is more spread out when there is noise (see previous question), which means that the different hypotheses are closer together in terms of their likelihood (see figure from previous question). As a consequence, the posterior probabilities of the different hypotheses will also be more similar (assuming the prior distribution doesn't change when we move from the regular to the noisy scenario). If the posteriors are close together, then the different hypotheses will be harder to tell apart.