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 word segmentation. The goal of this tutorial is to revise what you have learned by performing practical exercises.

1.1 Statistical Regularities

> *General guidance:* Conceptually, the important idea here is conditional probability (what does it mean intuitively, and why is $P(x|y)$ not the same as $P(y|x)$?). Explain that transitional probability is a type of conditional probability. Also, try to build an intuition of why transitional probability is informative for word segmentation (maybe be mentioning other indicators of word boundaries that were covered in the lecture). Then explain how probabilities can be estimated from frequencies, and take students through the relevant computations. In terms of terminology, please explain joint and marginal frequency as well (perhaps using Table 2).

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 learned 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., $P(t|' ')$. 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 colored in gray 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).

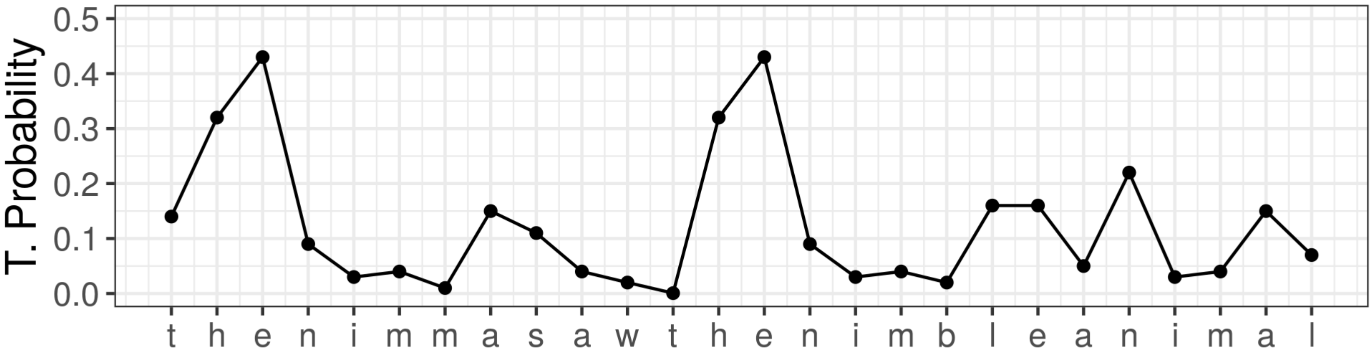

Figure 1: Transitional probabilities for the sequence *then|im|masaw|then|im|ble|an|imal*.

1.2 Minimum Description Length

General guidance: Here, an important idea to get across is that transitional probability (or other ways of guessing word boundaries) is not sufficient by itself, because we have a whole utterance to segment, and there are a lot of segmentation possibilities – so we need a way of adjudicating between them. And MDL provides such a way. For MDL itself, the key concept is that it formalizes the trade-off between an efficient lexicon and an efficient encoding of the input sequence (the utterance to be segmented). Maybe also point out that these two things don't have to be equally weighted, this is just a convenient assumption we're making here.

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 favored by the MDL model? ⇒ MDL minimizes the length of words, as shorter words are more plausible → favors segmentation 1. But MDL minimizes the number of different words (types) and maximizes the probability of each word (prefers fewer words) → favors segmentation 2.

2. Compute the MDL for the two segmentation hypotheses. Which hypothesis is favored by the MDL model?
⇒ See below. Segmentation 2 is favored 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	1 the 2 nimma 3 saw 4 nimble
7 al 8 cute	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