Informatics 1 CG – Tutorial 3

Carina Silberer

Week 4

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

1 Word Segmentation: Statistical Regularities

Background In class, you discussed 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:

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

where the frequency of occurrence of both events x and y, divided by the frequency of event x.

Exercise 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. 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., $frequ(a,m) = 245$. 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).

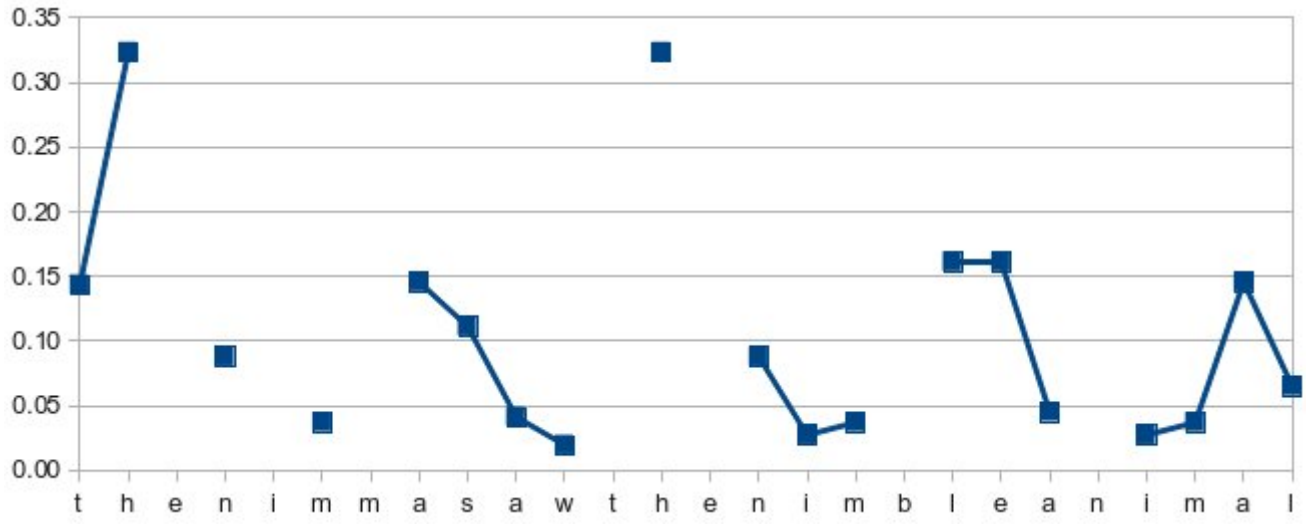

Figure 1: Transitional probabilities for the sequence *thenimmasawthenimbleanimal*.

2 Word Segmentation: Minimum Description Length

Exercise 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?
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

3 Lexicon Learning

In the lectures you talked about the first task involved in word learning, namely learning to segment a stream of sounds into words. You also discussed the second task that consists of learning to pair sounds with meanings.

Exercise Design a model that maps words to their meaning, i.e., the object a word refers to. Table 3 gives a lexicon of word-object mappings[1] the model should ideally learn. Of course your model needs some data from which it can learn the mapping. This data is a set of situations (in which a mother talks to a child). Each situation consists of an utterance and objects that are present and visible. Some examples are given in the Table 4

Hint You can design a model that makes use of statistics.

- You can use conditional probability, where you interpret x or y as word or object, respectively.
- Or measure the *association frequency* between a word and an object:

$$P(word, object) = \frac{frequ(word, object)}{frequ(word_i) + frequ(object_j)} \tag{3}$$

[1] source: `http://www.stanford.edu/~mcfrank/materials/ww_model/data/`

Word	Object	Word	Object
baby	BABY	bear	BEAR
bigbird	BIRD	bigbirds	BIRD
bird	BIRD	book	BOOK
books	BOOK	bunny	BUNNY
bunnyrabbit	BUNNY	cow	COW
cows	COW	moocow	COW
moocows	COW	duck	DUCK
duckie	DUCK	eyes	EYES
hand	HAND	hat	HAT
kitty	KITTY	kittycat	KITTY
kittycats	KITTY	lamb	LAMB
lambie	LAMB	mirror	MIRROR
pig	PIG	piggie	PIG
piggies	PIG	rattle	RATTLE
ring	RING	rings	RING
sheep	SHEEP	bunnies	BUNNY
birdie	DUCK	bird	DUCK

Table 3: Lexicon of word-object mappings.

Utterance	Objects
ahhah look we can read books david	BOOK BIRD RATTLE FACE
thats a nice book	BOOK BIRD RATTLE KITTY BABY
the bear has a baby bottle	BOOK BIRD RATTLE FACE BEAR
yes david has baby bottles	BOOK BIRD RATTLE FACE BEAR
and a bear with a bottle	BOOK EYES BEAR
theres a mirror	BOOK BIRD RATTLE MIRROR BUNNY
does david want to read the book	BOOK EYES
ah a bunny	BOOK BIRD RATTLE MIRROR BUNNY
what do bunnies do	BOOK BIRD RATTLE MIRROR BUNNY
bunnies go hiphop hiphop	BOOK BIRD RATTLE MIRROR BUNNY
lots of toys	RING HAND
we watch big bird dont we	BIRD RATTLE

Table 4: Example of situations: pairs of utterances and objects present

.