Statistical and Adaptive Natural Language Processing

T-61.5020 (5 cr) L, Spring 2008

Seventh lecture

Contextual Information and Compositionality

Lecturer: **Jaakko Väyrynen**

Slides: Krista Lagus, Mathias Creutz, Timo Honkela

Lecture based on:

- Chapter 5 in Manning & Schütze.
- Mathias Creutz. 'Induction of the morphology of natural language: unsupervised morpheme segmentation with application to automatic speech recognition', PhD Thesis, 2006.

7. Contextual Information and Compositionality

- The importance of **context** in the interpretation of language: for instance ambiguous word meanings:
  - "Aloitin alusta."
  - "Alusta kovalevy!"
  - "Näin monta alusta."
- **Compositionality** of meaning and form (within and beyond words):
  - "I saw you yesterday." (= I + SEE + YOU + YESTERDAY)
  - "openmindedness" (= OPEN + MIND + -ED + -NESS)
  - "What is chewing gum made of?"
  - "The New Scientist reports that people who chewed gum during the memory tests scored higher than those who did not."
  - "How do you do."

7.1 Word Context: “Meaning Is Use”

- What is meaning? What is meaning in language?
  - cognitive linguistics (1970 –) vs. structuralists (18xx – 1970)
- Words are not labels for real things in the world. They refer to ideas that we have of the world. (Saussure)
- To understand a particular word, one has to know in which possible combinations with other words it occurs. Consider less obvious cases such as *time, consider, the, of*. (Harris)
- You shall know a word by the company it keeps. (Firth)
- Meaning is use. (Wittgenstein)

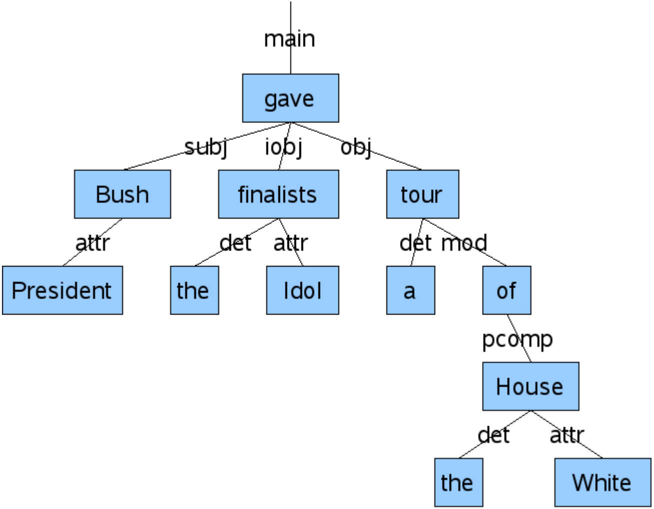
⇒ *Computational approach:* The typical context in which a word occurs can be used to characterize the word in relation to other words.

Types of Contexts

- n-grams, dynamical context
  - *I don't like . . .*
- fixed window, word-document matrix
  - $\textit{I}\ \boxed{\textit{don't like } \boxed{\textit{icecream}}\textit{, but I}}$ *do like cookies.*
  - Represent words as vectors
- bag of words vs. position sensitivity
  - Are different positions in the word context mapped onto different subvectors?
- distant dependencies and phrasal structure:

Example: Phrasal Structure Using Functional Dependency Grammar (FDG)

- President Bush gave the Idol finalists a tour of the White House.



7.2 Expressions with Limited Compositionality: Collocations

- A *collocation* is a conventional phrase that consists of one or more words
- Examples:
  - 'weapons of mass destruction', 'disk drive', 'part of speech' (as compound words in Finnish 'joukkotuhoaseet', 'levyasema', 'sanaluokkatieto')
  - 'bacon and eggs'
  - verb selection: 'make a decision' not 'take a decision'.
  - adjective selection: 'strong tea' but not 'powerful tea'; 'vahvaa teetä', rarely 'voimakasta teetä' (the choices may reflect cultural positions: strong → {tea, coffee, cigarettes}, powerful → {drugs, antidote})
  - 'kick the bucket', 'heittää veivinsä' (euphemism, saying, idiom)

- Names that individualize beings, societies, or events: 'White House' Valkoinen talo, 'Tarja Halonen'

- Concepts that overlap with collocations: term, technical term, terminological phrase. NB. in information retrieval 'a term' has a broader meaning: 'word or collocation'.

Word frequence and part-of-speech filtering

Only frequence:

Example: Is it more natural to say “strong tea” or “powerful tea”?
Solution: Search with Google: “strong tea” 289 000; “powerful tea” 3 420

Sufficient method for some specific questions. However, if bigrams are ordered according to frequency, the best ones are 'of the', 'in the', 'to the', . . .

Frequence + part-of-speech (POS):

If the POS for each word is known, and 'allowed' POS patterns for collocations can be described:

- Order word pairs (or tuples) according to frequency (count)
- Accept only certain POS patterns:
  AN, NN, AAN, ANN, NAN, NNN, NPN (Justeson & Katz's POS filter)

$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

Word distance mean and variance

How about more flexible collocations, that contain words in the middle that are not part of the collocation?

Calculate mean and variance for distance. If mean is not zero and variance is small, it is a potential collocation (NB. assuming distance is normally distributed).

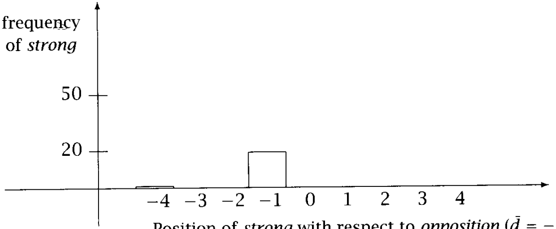
E.g. '*knock . . . door*' (not 'hit', 'beat', or 'rap'):
a) 'she *knocked* on his *door*'
b) 'they *knocked* at the *door*'
c) '100 women *knocked* on Donaldson's *door*'
d) 'a man *knocked* on the metal front *door*'

Algorithm

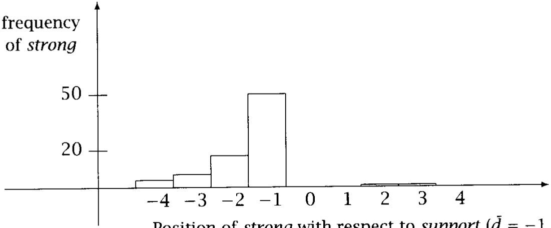
- Slide a fixed size window over the text (e.g., width 9) and collect all word pair instances in the whole text

- Count mean of word distances:
  $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{4}(3+3+5+5) = 4.0$
  (if apostrophe and 's' are counted as words)

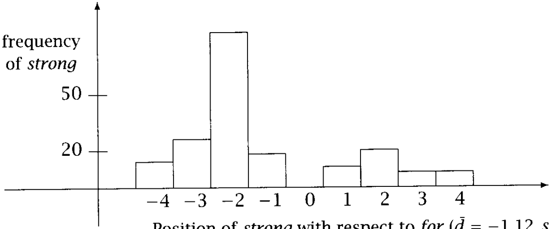
- Estimate variance s^2 (small sample sizes):
  $s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = \frac{1}{3}((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2)$
  $s = 1.15$



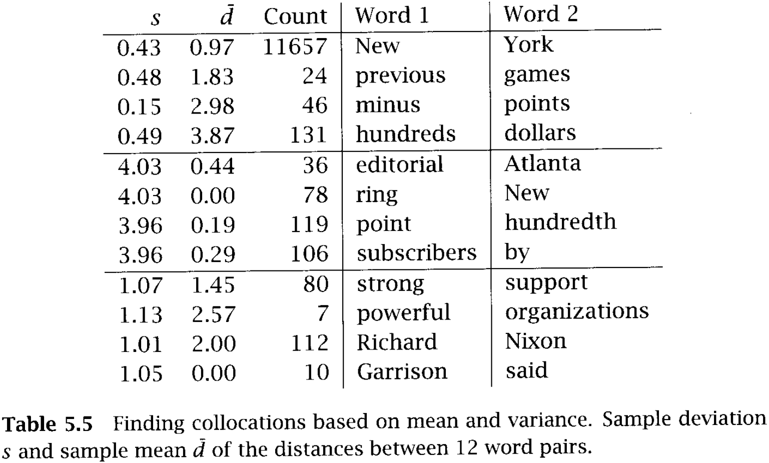
Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).



s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Table 5.5 Finding collocations based on mean and variance. Sample deviation s and sample mean $\bar{d}$ of the distances between 12 word pairs.

Consider:

1. What happens if the words have two or more typical positions related to each other?
2. What is the significance of the window width?

7.3 Hypothesis Testing (for the Discovery of Collocations)

Is the large hit count a coincidence (e.g., base frequence for one of the words is high)? Do two words occur more together than randomness would suggest?

1. Formulate *null hypothesis* H_0: the association is random
2. Count prob. p that words co-occur if H_0 is true
3. Abandon H_0 if p is too low, less than significance level, e.g., $p < 0.05$ or $p < 0.01$.

The definition of independence is applied to the null hypothesis.

Assuming that the word pair probability, if H_0 is true, is a product of the probabilities of each word:
$P(w^1 w^2) = P(w^1) P(w^2)$

The t Test

A statistical test for determining whether the sample mean differs from the expected mean of a generative data distribution. It assumes that probabilities are approximately normally distributed.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}, \text{ where} \tag{1}$$

$\bar{x}$, s^2: sample mean and variance, N: number of samples, and μ: mean of distribution. Choose wanted p level (0.05 or lower). Read the corresponding upper bound for t from the table. If t is higher, discard H_0.

Discovery of Collocations Using the t Test

The null hypothesis is that word co-occurrences are random: Example: H_0 : $P(new\ companies) = P(new)P(companies)$

$\mu = P(new)P(companies)$
$\bar{x} = \frac{c(new\ \ companies)}{c(\cdot,\cdot)} = \hat{p}$
$s^2 = p(1-p) = \hat{p}(1-\hat{p}) \approx \hat{p}$ (true for the Bernoulli distribution and low p)
$N = c(\cdot,\cdot)$

Bernoulli: special case of binomial distribution: $b(r;n,p) = \binom{n}{r}p^r(1-p)^{n-r}$, such that $n = 1$ and $r \in \{0,1\}$.

- Sort words in order according to the test OR

- Hypothesis testing: Choose significance level ($p = 0.05$ or $p = 0.01$) and see the value from the t table for which higher values mean discarding the null hypothesis.

Pearson's Chi-Square Test χ^2

- χ^2 test measures the dependence of variables based on the definition of independence: if two variables are independent, the joint distribution is the product of marginal distributions
- A distribution for two variables can be represented as a two-dimensional contingency table ($r \times c$).
- Count *for each table cell* (i, j) the difference between observed distribution O (real joint distribution) and the expected distribution E (product of marginal distributions), and take the sum scaled with the expected value of the distribution:

$$X^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{2}$$

  where $E_{i,j} = O_{i,\cdot} * O_{\cdot,j}$.

- X^2 is *asymptotically* χ^2 distributed. A problem still: sensitive to sparse data.

- Thumb of rule: do not apply the test if $N < 20$ or if $20 \leq N \leq 40$ and some $E_{i,j} \leq 5$

Discovery of Collocations Using the χ^2 Test

Formulate the problem so that for each word there is one random variable that can have two values (the word either occurs or doesn't occur in a word pair instance).

The joint distribution for words can then be represented as a 2×2 table. E.g.

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8	4 667
$w_2 \neq companies$	15 820	14 287 181

In the case of a 2×2 table, Eq. 2 has the form:

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

- Sort words in order according to the test OR

- Hypothesis testing: Choose significance level ($p = 0.05$ or $p = 0.01$)

and see the value from the χ^2 table for which higher values mean discarding the null hypothesis.

Problem with collocation identification

This approach doesn't separate negative and positive dependencies. I.e., if the words shun each other, the test will also give a high value, because then there indeed is a dependency between the occurrences of the words. In collocation detection, however, only positive dependencies are interesting.

Other applications for the χ^2 test:

- Machine translation: Identification of word translation pairs from sentence aligned corpus (cow, vache: co-occurrences are due to dependency).
- A similarity metric between two corpora: $n \times 2$ table in which for each observed word w_i, $i \in (1 \ldots n)$, the frequency of the word in question in corpus j is reported.

Likelihood Ratios

How much more likely is H_2 than H_1? Calculate the likelihood ratio λ:

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}$$

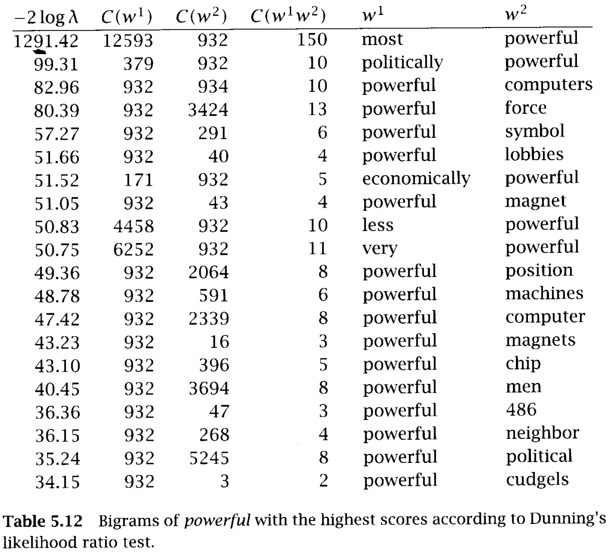
Example:
H_1: w_1 and w_2 are independent: $P(w_2|w_1) = p = P(w_2|\neg w_1)$
H_2: w_1 and w_2 are dependent: $P(w_2|w_1) = p1 \neq p2 = P(w_2|\neg w_1)$

Assume clear positive dependency, i.e., $p_1 \gg p_2$.

Apply ML estimates (means) while counting p, p_1 and p_2:
$p = \frac{c_2}{N}$, $p_1 = \frac{c_{12}}{c_1}$, $p_2 = \frac{c_2 - c_{12}}{N - c_1}$

Assume binomial distributions. E.g., $p(w_2|w_1) = b(c_{12}; c_1, p)$. Describe the simultaneously effective constraints for each model as a product. Final result: Eq. 5.10 in the book.

$-2 \log \lambda$ is asymptotically χ^2 distributed. It has also been shown that for sparse data the likelihood ratio gives a better approximation for the χ^2 distribution than the X^2 statistic.



$-2\log\lambda$	$C(w^1)$	$C(w^2)$	$C(w^1w^2)$	w^1	w^2
1291.42	12593	932	150	most	powerful
99.31	379	932	10	politically	powerful
82.96	932	934	10	powerful	computers
80.39	932	3424	13	powerful	force
57.27	932	291	6	powerful	symbol
51.66	932	40	4	powerful	lobbies
51.52	171	932	5	economically	powerful
51.05	932	43	4	powerful	magnet
50.83	4458	932	10	less	powerful
50.75	6252	932	11	very	powerful
49.36	932	2064	8	powerful	position
48.78	932	591	6	powerful	machines
47.42	932	2339	8	powerful	computer
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

Table 5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test.

Pointwise Mutual Information

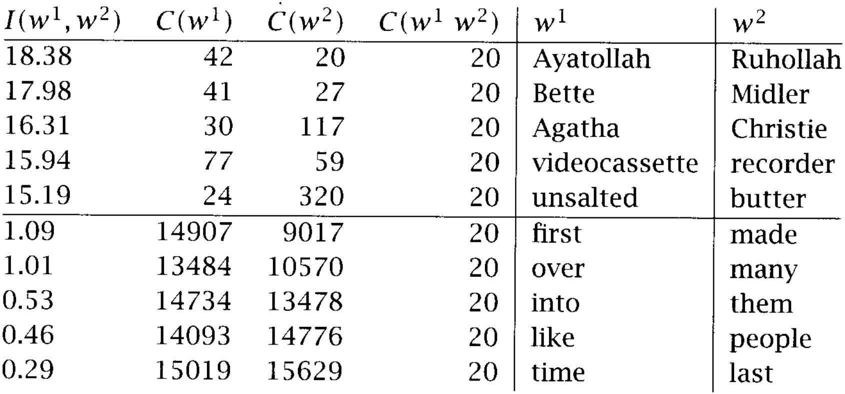
Lets recall the formulas for entropy $H(x)$ and mutual information $I(x;y)$:

$$\begin{aligned} H(x) &= -E(\log p(x)) \\ I(X;Y) &= H(Y) - H(Y|X) = (H(X) + H(Y)) - H(X,Y) \\ &= E_{X,Y}(\log \tfrac{p(X,Y)}{p(X)p(Y)}) \end{aligned}$$

which represents the *average* information that both x and y contain.

Lets define *pointwise mutual information* between some specific events x and y (Fano, 1961):
$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$

Could it be applied to selecting collocations? Motivated by intuition: if there is high mutual information between words (i.e., the information communicated by both words is high), it may assumed to be a collocation.



$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

Table 5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.

We can see from Table 5.16 that if either of the words is infrequent, the score is high.

Mutual information for completely dependent words:

$$I(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x)}{p(x)p(y)} = \log \frac{1}{P(y)}$$

grows when the words get less frequent. Extreme: two words both occur only once and at that time together. However, there is little evidence of a collocation which is lost.

Conclusion: Not too good score for this purpose, misleading with especially low probabilities. As a consequence data sparseness is very badly tolerated.

I_{1000}	w^1	w^2	w^1w^2	Bigram	I_{23000}	w^1	w^2	w^1w^2	Bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Table 5.16 Problems for Mutual Information from data sparseness. The table shows ten bigrams that occurred once in the first 1000 documents in the reference corpus ranked according to mutual information score in the first 1000 documents (left half of the table) and ranked according to mutual information score in the entire corpus (right half of the table). These examples illustrate that a large proportion of bigrams are not well characterized by corpus data (even for large corpora) and that mutual information is particularly sensitive to estimates that are inaccurate due to sparseness.

Summary: Hypothesis Testing

My guess is that you will not use hypothesis testing in order to discover collocations (except in the exercise session and possibly in the exam of this course).

However, you will need hypothesis testing if you do research and compare the performance of your system to some other system. You will need to prove that your system performs better than your competitor in a *statistically significant way*.

- T-test (assumes Gaussian distribution)
- Pearson's chi-square test (small sample sizes are not sufficient)

Also take a look at (not taught in this course):

- Sign test
- Wilcoxon signed-rank test

7.4 Compositionality Within Words

Just as it may be important to recognize that the vocabulary of a language partly consists of multi-word expressions (collocations), it is important to recognize that words have inner structure and are related to one another.

Linguistic *morphology* studies how words are formed from *morphemes*, which are the minimal meaningful form-units. Morphemes are portions of utterances that recur in other utterances with approximately the same meaning.

For instance, the Finnish morpheme *talo* (house):

aave+talo+i+sta, aika+kaus+lehti+talo+j+en, aika+talo+a, aikuis+koulu+t+us+talo+n, aikuis+talo+uks+i+en, aito+talo+ude+n, akatemia+talo, akatemia+talo+n, akatemia+talo+ssa, akvaario+talo+us, alas+talo, alas+talo+n, alas+talo+on, ala+talo, ala+talo+a, ala+talo+kin, ala+talo+lla, ala+talo+n, ala+talo+on, ala+talo+sta, . . .

Zellig Harris' Heuristic Morpheme Segmentation Method (1955)

Morpheme boundaries are proposed at intra-word locations with a *peak in successor and predecessor variety*.

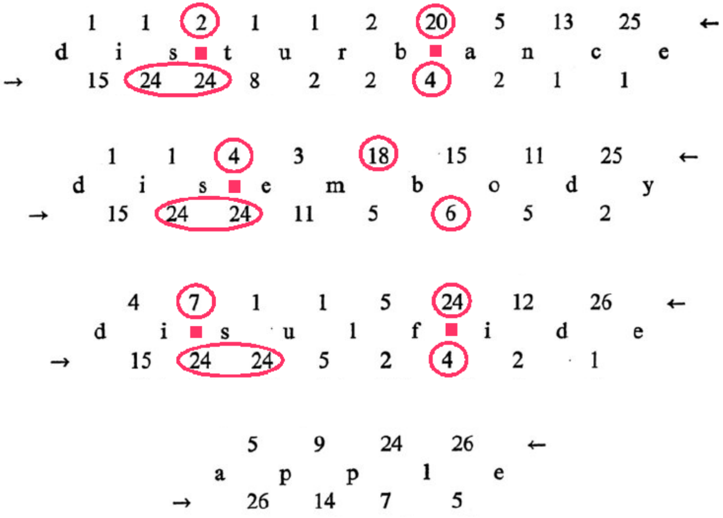
Let us take a look at some examples of successor variety:

- How many different letters can continue an English word starting with 'd'? (d**a**y, d**e**bt, d**i**g, d**o**g, d**r**ill, ...)
- How many different letters can continue a word starting with 'drea'? (drea**m**, drea**d**ful, ...)
- How many different letters can continue a word starting with 'dream'? (dream**s**, dream**y**, dream**i**ly, dream**b**oat, ...)

Predecessor variety works in the same way, but the words are read backwards:

- How many different letters can precede 'ily' at the end of an English word? (drea**m**ily, hap**p**ily, fun**n**ily, ...)

Example of Harris' Method



Note that there are more refined versions, e.g., Hafer and Weiss (1974).

Statistical Approaches (1995–)

Think of it as a *compression problem*. If we observe a corpus of natural language containing lots of different word forms, what kind of model could explain the kind of data we observe in an elegant manner?

From what kind of *lexicon* could words such as apple, orange, lemon, juice, applejuice, orangejuice, appletree, lemontree emerge?

Suppose that lexicon candidates are created using a statistical generative process: characters are drawn from an alphabet including a *morph break* character:

Lexicon 1 “a␣c␣e␣g␣i␣j␣l␣m␣n␣o␣p␣r␣t␣u␣␣” (14 morphs),
$P(\textit{Lexicon 1}) = (\frac{1}{27})^{29}$

Lexicon 2 “apple␣juice␣lemon␣orange␣tree␣␣” (5 morphs),
$P(\textit{Lexicon 2}) = (\frac{1}{27})^{31}$

Lexicon 3 “apple␣applejuice␣appletree␣juice␣lemon␣lemontree␣orange␣orangejuice␣␣” (8 morphs), $P(\textit{Lexicon 3}) = (\frac{1}{27})^{69}$

Statistical Approaches: Continued

Rewrite the words in the corpus (data) as morphs and compute probability (# is a word boundary morph):

Data | Lexicon 1 "`a p p l e # o r a n g e # l e m o n # j u i c e # a p p l e j u i c e # o r a n g e j u i c e # a p p l e t r e e # l e m o n t r e e # #`" The sequence consists of 69 morphs and its probability conditioned on *Lexicon 1* is: $P(\textit{Data}\,|\,\textit{Lexicon 1}) = (\frac{1}{14+1})^{69} \approx 7.1 \cdot 10^{-82}$.

Data | Lexicon 2 "`apple # orange # lemon # juice # apple juice # orange juice # apple tree # lemon tree # #`". The sequence consist of 21 morphs, and $P(\textit{Data}\,|\,\textit{Lexicon 2}) = (\frac{1}{5+1})^{21} \approx 4.6 \cdot 10^{-17}$.

Data | Lexicon 3 "`apple # orange # lemon # juice # applejuice # orangejuice # appletree # lemontree # #`". The sequence consists of 17 morphs, and $P(\textit{Data}\,|\,\textit{Lexicon 3}) = (\frac{1}{8+1})^{17} \approx 6.0 \cdot 10^{-17}$.

Statistical Approaches: Maximum A Posteriori (MAP) Optimization

The model selection procedure is based on maximizing the posterior probability of the model $P(\textit{Lexicon X} \,|\, \textit{Data})$. The posterior can be rewritten using the Bayes' rule:

$$\begin{aligned} P(\textit{Lexicon X} \,|\, \textit{Data}) &= \frac{P(\textit{Lexicon X}) \cdot P(\textit{Data} \,|\, \textit{Lexicon X})}{P(\textit{Data})} \\ &\propto P(\textit{Lexicon X}) \cdot P(\textit{Data} \,|\, \textit{Lexicon X}). \end{aligned}$$

$P(\textit{Lexicon 2}) \cdot P(\textit{Data} \,|\, \textit{Lexicon 2}) \approx 1.9 \cdot 10^{-61} > P(\textit{Lexicon 3}) \cdot P(\textit{Data} \,|\, \textit{Lexicon 3}) \approx 1.0 \cdot 10^{-115} > P(\textit{Lexicon 1}) \cdot P(\textit{Data} \,|\, \textit{Lexicon 1}) \approx 2.2 \cdot 10^{-123}$.

In this comparison the complexity of the model has been balanced against the fit of the training data, which favors a good compromise, that is, a model that does not overlearn and that adequately generalizes to unseen data.

Statistical Approaches: Demo

Variations of the approach sketched out above have been used for both *word segmentation* (Asian languages, where words are written without explicit boundaries) and *morphology modeling*, e.g.,

- Carl de Marcken (1995)
- Michael R. Brent (1999)
- John Goldsmith (2001) (*Linguistica*)
- Mathias Creutz and Krista Lagus (2002–) (*Morfessor*)
- Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson (2006)

Morfessor demo: http://www.cis.hut.fi/projects/morpho/