Spellcheck

Acknowledgements : Most slides are based on : J&M's Speech and NLP. I Other sources are acknowledged separately in individual slides.

Two key problems

- Problem 1: Words have a grammar, just as sentences do. How do we map variants of words to their root forms?

- Problem 2: How do we overcome spelling errors?

Corresponding to each of these problems, there is Science + Engineering

Some questions

- Why is correcting spelling errors important?
- Why are spelling errors made?
- What are the categories of spelling errors?
- What factors can we exploit to recover from them?

Some questions

- Why is correcting spelling errors important?
  - Human errors in typing (search engines)
    - How rampant are spelling errors?
  - OCR
  - Speech-to-text
- Why are spelling errors made?
  - Homophones
  - No neat mapping between structure and pronunciation of words

Britney Spears

Categories of spelling errors

- Non-words
  - *Seperate* for *separate*
- Words
  - *Dessert* for *desert*
  - *Piece* for *peace*

An important question: Is context available?

Categories of spelling errors

- Non-words
  - *Seperate* for *separate*
- Words
  - *Dessert* for *desert*
  - *Piece* for *peace*

  An important question: Is context available?

An alternative classification:

-- Typographic errors (homologous errors because of keyboard)

-- Cognitive errors

How can we recover from spelling errors?

- Edit distance
- Keyboard
- Pronunciation
- Context
- Syntax
- Source-specific issues
  - OCR : cl and d

Two classes of systems

- Spelling error detection
- Spelling error correction

Typing errors

- Single error mis-spellings
  - Insertion
  - Deletion
  - Substitution
  - Transposition

OCR errors

- Correct:

  The quick brown fox jumps over the lazy dog.
- Recognized:

  ‘lhe q~ick brown foxjurnps ovcr tb l azy dog.

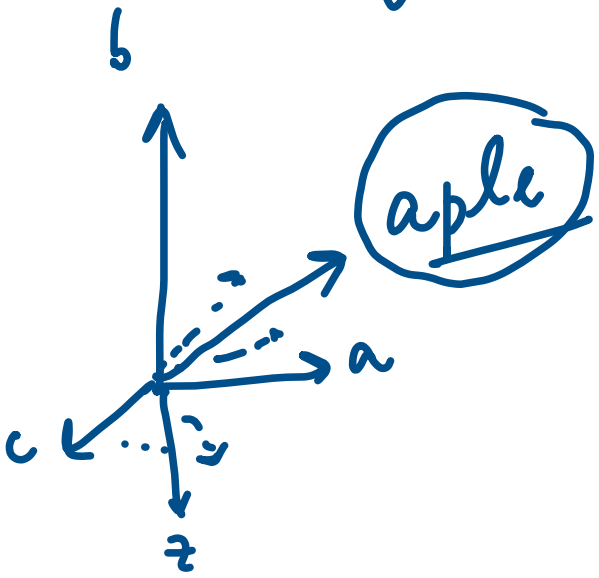
  – Substitutions
  – Multisubstitutions (framing errors)
  – Space deletions
  – Space insertions
  – Failures

Other issues...

- You may be right, the document may be wrong...
- Google may be unfair to newcomers...

Spelling correction.

aple



apple
able
maple
ape
⋮

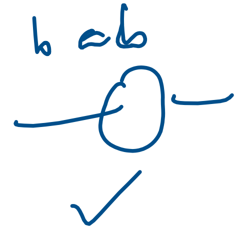
a b c ... l ... p ... z

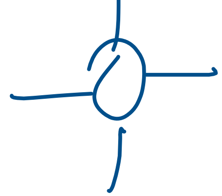
[1 0 0 1 .. 1 .. 1, 0 . 0 0 ..]

order

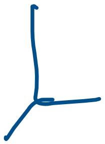
[1 0 0 0 4 0 0 3 0 2 ...]

Order :





bcacb

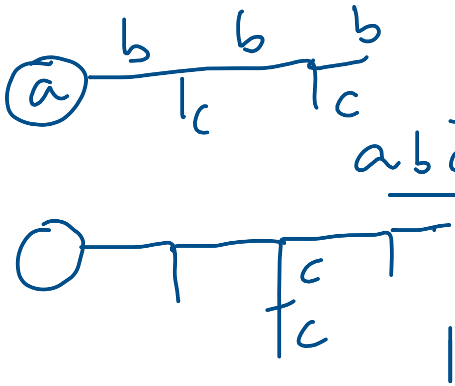


ML → Decision Theoretic ML

ML → Syntactic ML

grammar induction.

abcbcb



abcbccbcb

Keys

abccbcbcbb

O – a

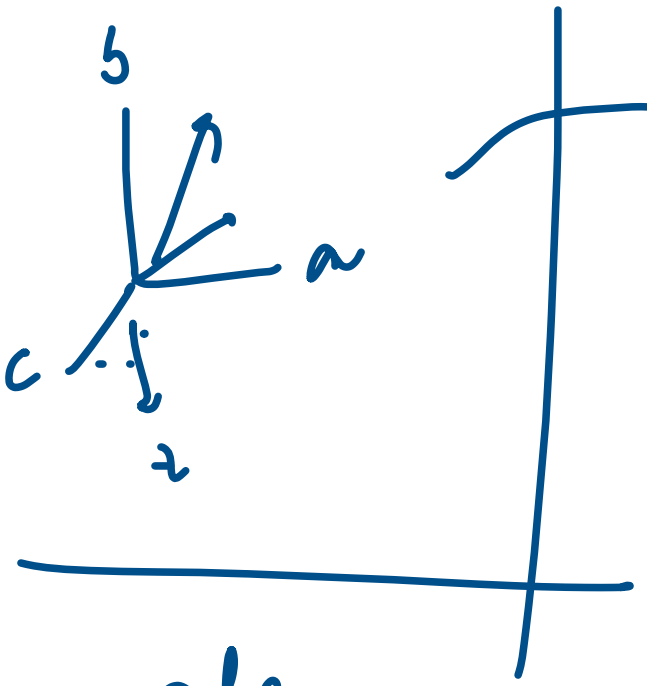
⊣ – b

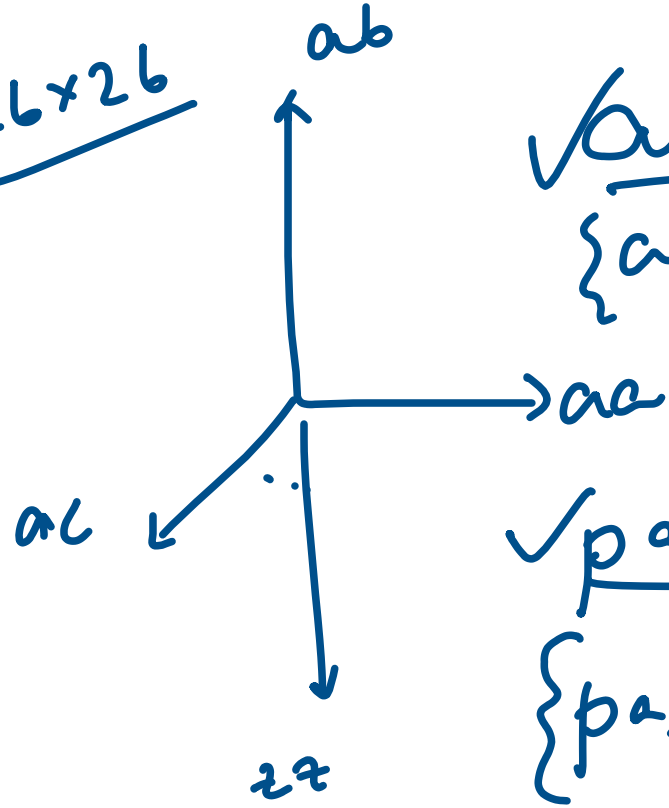
I – c

Keys / non-Keys

$a(b^+c^+)^+ \; b^+$

Syntactic ML





apa

{ap, pa}

pap

{pa, ap}

aple

able

1, 2,

{ap, pl, le}

26×26

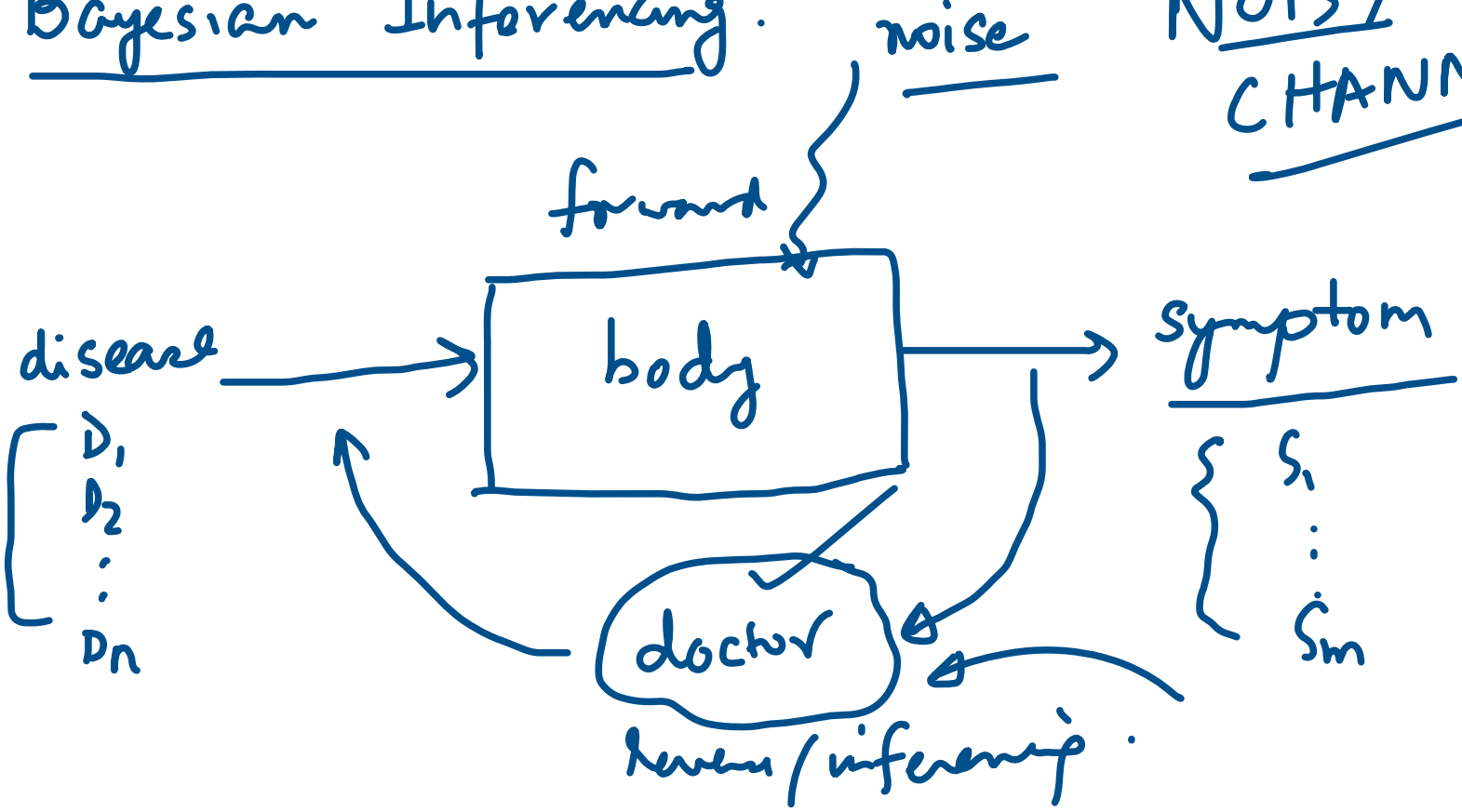
ap pl

aa ----- zz.

[0 0 .. 1 0 1 ..]

Bayesian Inferencing.

NOISY CHANNEL



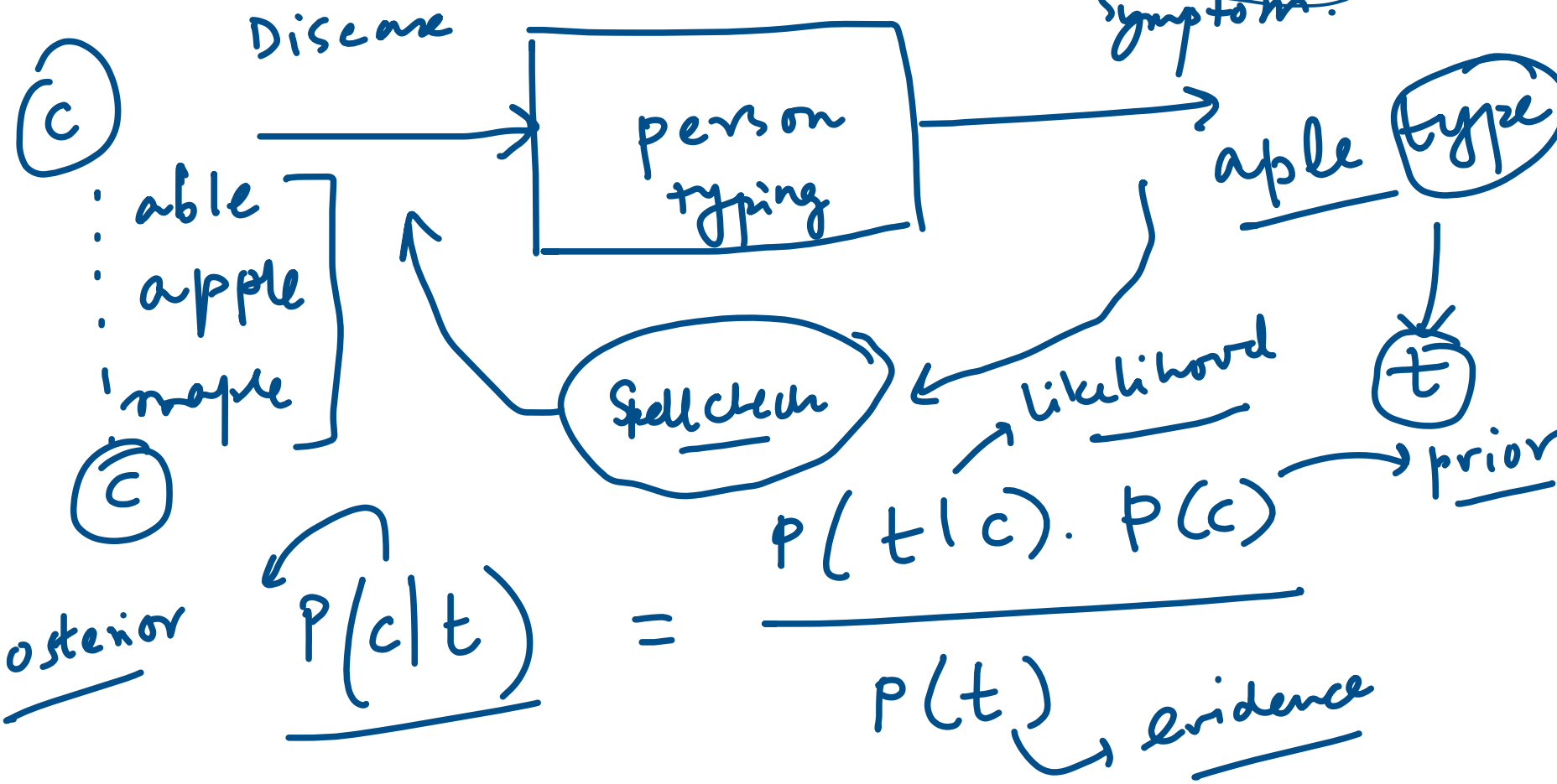
diagnosis. ✓

Spellcheck as Diagnosis

MBD

CONSISTENCY BASED DIAGNOSIS

ATE



typo

$$P(c|t) = \frac{P(t|c) \cdot P(c)}{P(t)}$$

L P

↑ correct

E

generative

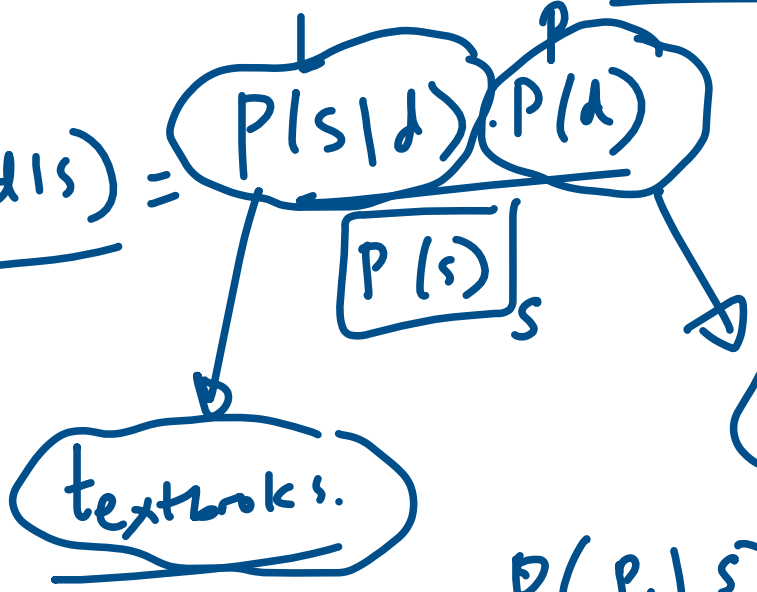
discriminative

Likelihood → Generative process

$P(d|s) = P(s|d) \cdot P(d)$

$P(s)$

s



$P(d_1|s)$

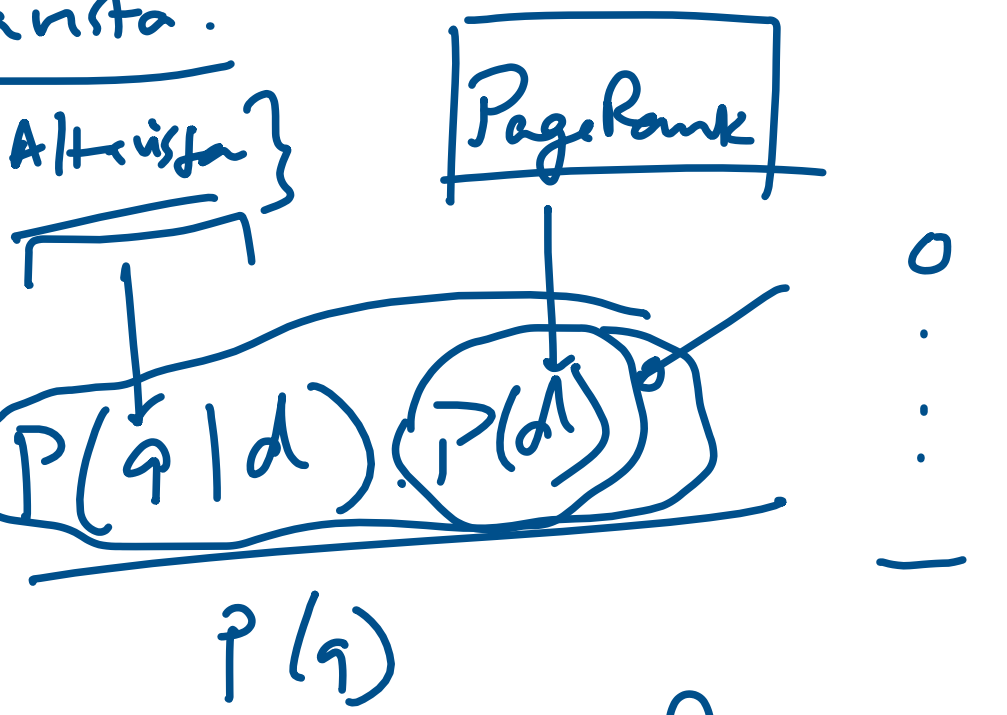
$P(d_2|s)$

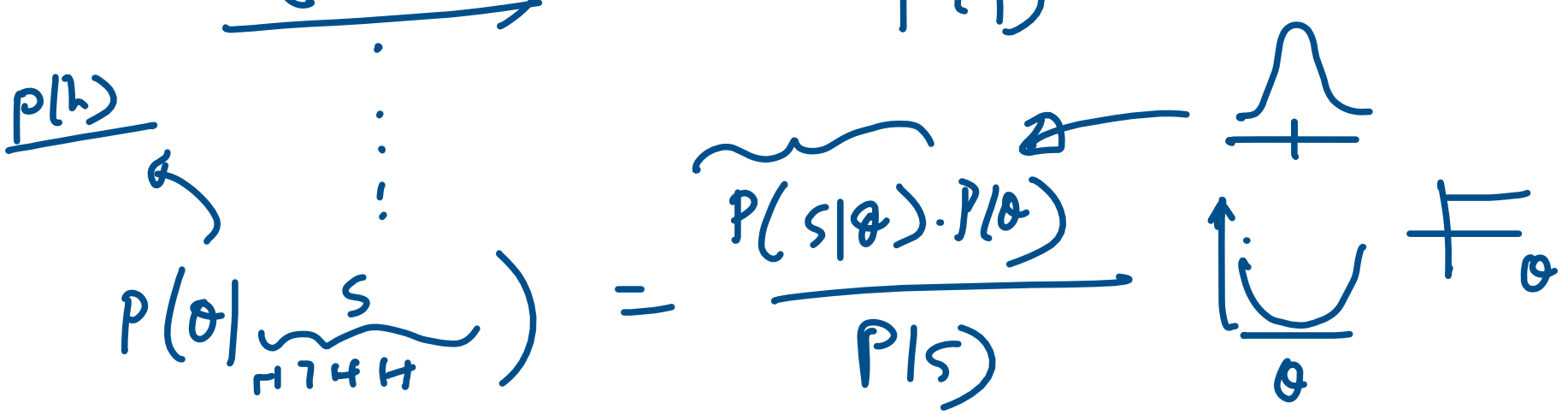
$d(d_n|s)$

$P(P_i|s) \propto P(s|P_i) \cdot P(P_i)$

Google vs. Altavista.

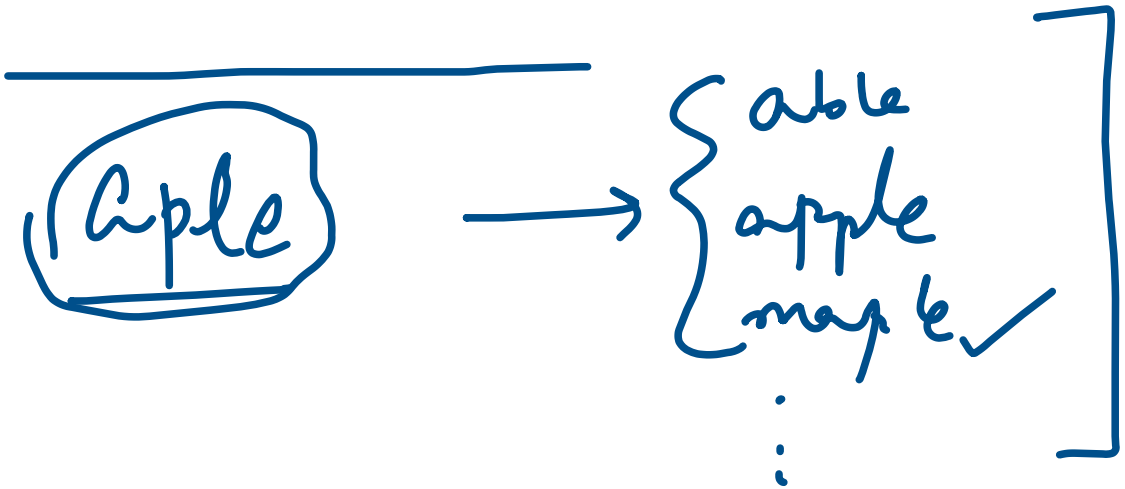
PageRank.





Subproblems.

- generate candidates.
- prior estimation
- likelihood estimation



bigrams

apple

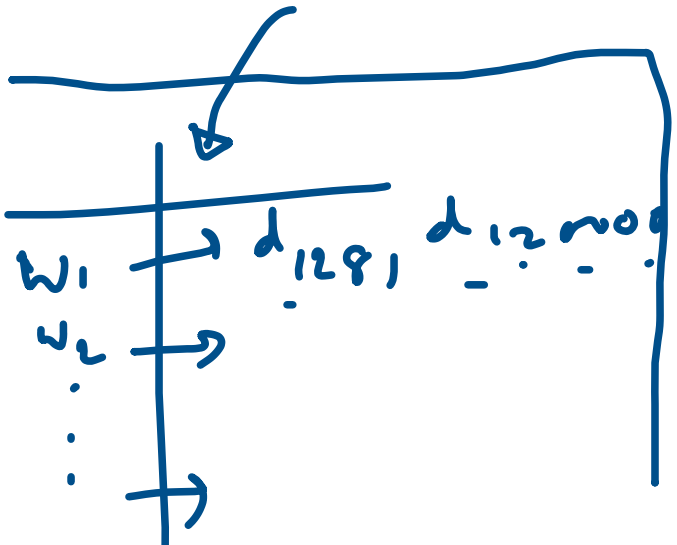
{ap, pl, le}

postings file / inverted index

aa →
ab → [abbb,]
.
.
.
.
.
.
zz

26x26

zz



Prior

$$P(c|t) \propto P(t|c) \cdot P(c)$$

prior.

$\binom{m}{n}$

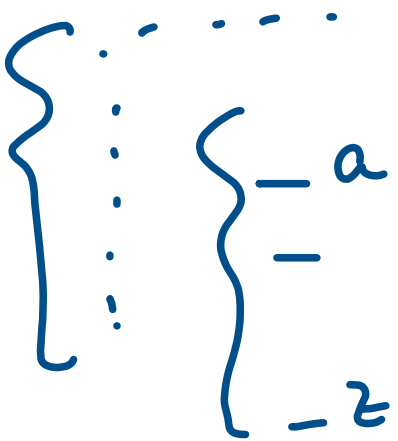
$n \to \infty$

—

the

obvious

corpus.



bottom-up.

$$P(c) = \frac{n}{N}$$

N:

al → {_a, al, l_}

{_at_} {_a, at, t_}

types : no. of distinct words

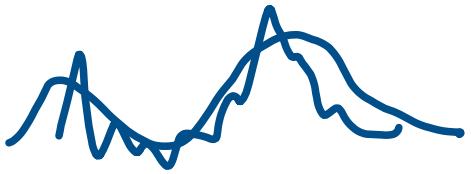
tokens : no. of words incl. repetitions.

"the man chased the monkey"

$V = 4$

$N = 5$

Smoothing.

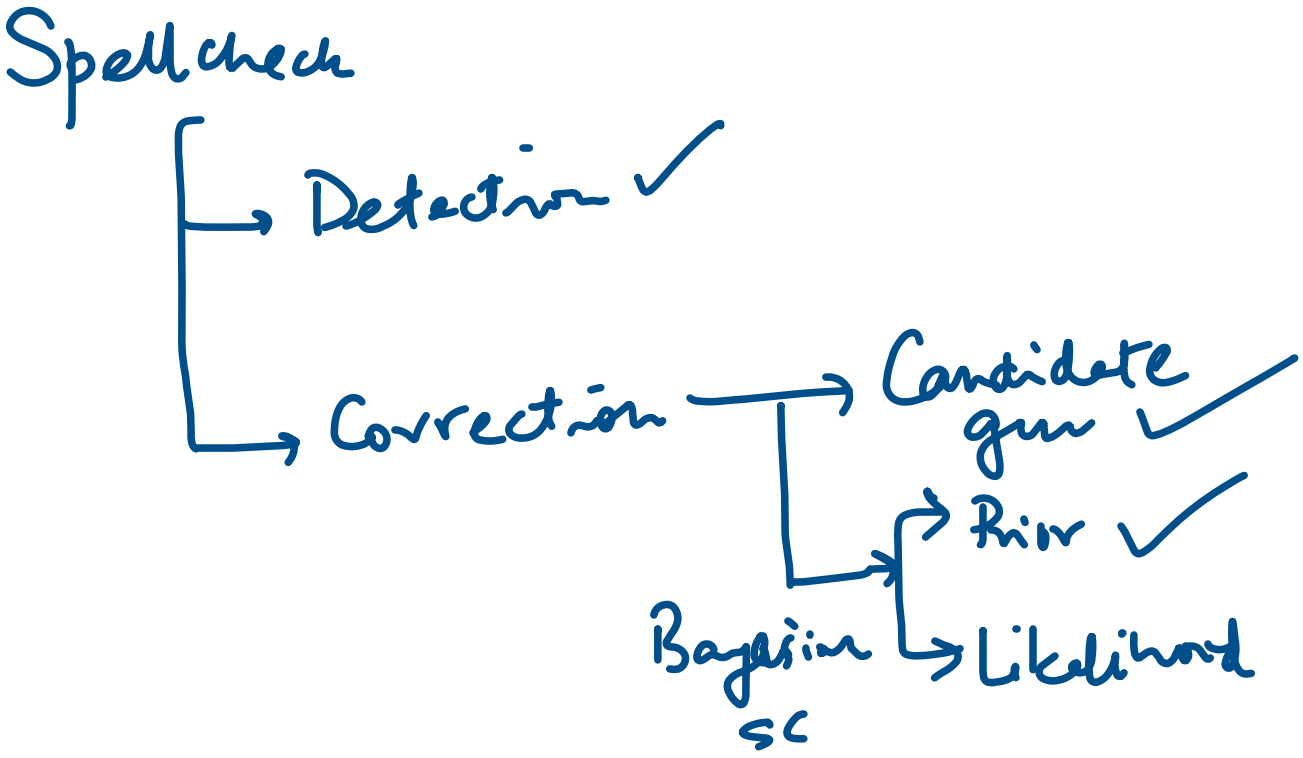


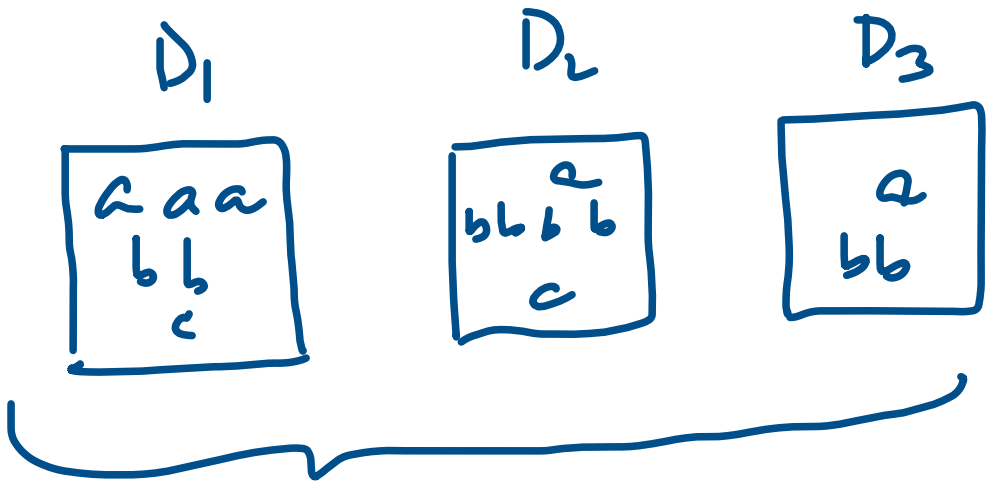
$$P(c) = \frac{n_c}{N}$$

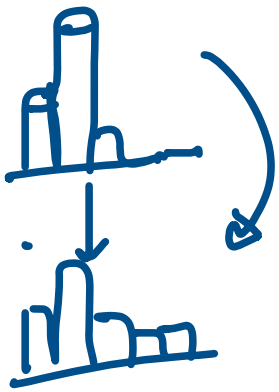
after smoothing

$$\frac{n_c + 1}{N + V}$$

Revised estimate of prior after smoothing


Spellcheck
Detection ✓
Correction
Candidate gen ✓
Prior ✓
Bayesian
Likelihood
sc





Laplace (add-1 smoothing)

B. $\frac{n_c}{N}$ $\left(\frac{n}{N}\right)$

A. $\left(\frac{n+1}{N+V}\right)$ a: $\frac{5}{15} \rightarrow \frac{6}{20}$

5 types

$\frac{n}{N} > \frac{n+1}{N+V}$

$nV > N$

$n > \frac{N}{V}$

$N = 15$

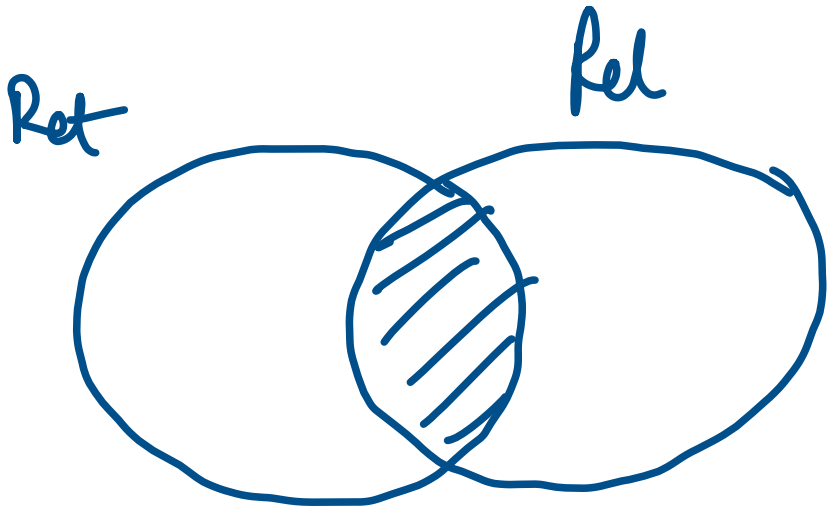
$V = 5$

$n > 3$

N:1 tokens

	B.S.	A.S.
a	5	6
b	8	9
c	2	3
d	0	1
e	0	1
	15	20

V = 5



Ret : Set of retrieved suggestions.

Rel : Set of relevant suggestions.

$$\text{Precision} : \frac{|Ret \cap Rel|}{|Ret|}$$

$$\text{Recall} : \frac{|Ret \cap Rel|}{|Rel|}$$

Candidate corrections

t : acress ✓

Error	Correction	Transformation: Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	–	2	deletion ✓
acress	cress	–	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	–	2	5	insertion
acress	acres	–	2	4	insertion

c → t

Prior probabilities

$$\hat{c} = \underset{c \in V}{\operatorname{argmax}} \underbrace{P(t|c)}_{L} \cdot \underbrace{P(c)}_{P}$$

cress → acress

c	freq(c)	p(c)
actress	1343	.0000315
cress	0 →	.000000014
caress	4	.0000001
access	2280	.000058
across	8436	.00019
acres	2879	.000065

types

Assoc. Press (88) ... 44 million words

Likelihood.

$$\frac{\text{Count}[pb]}{\text{Count}[p]}$$

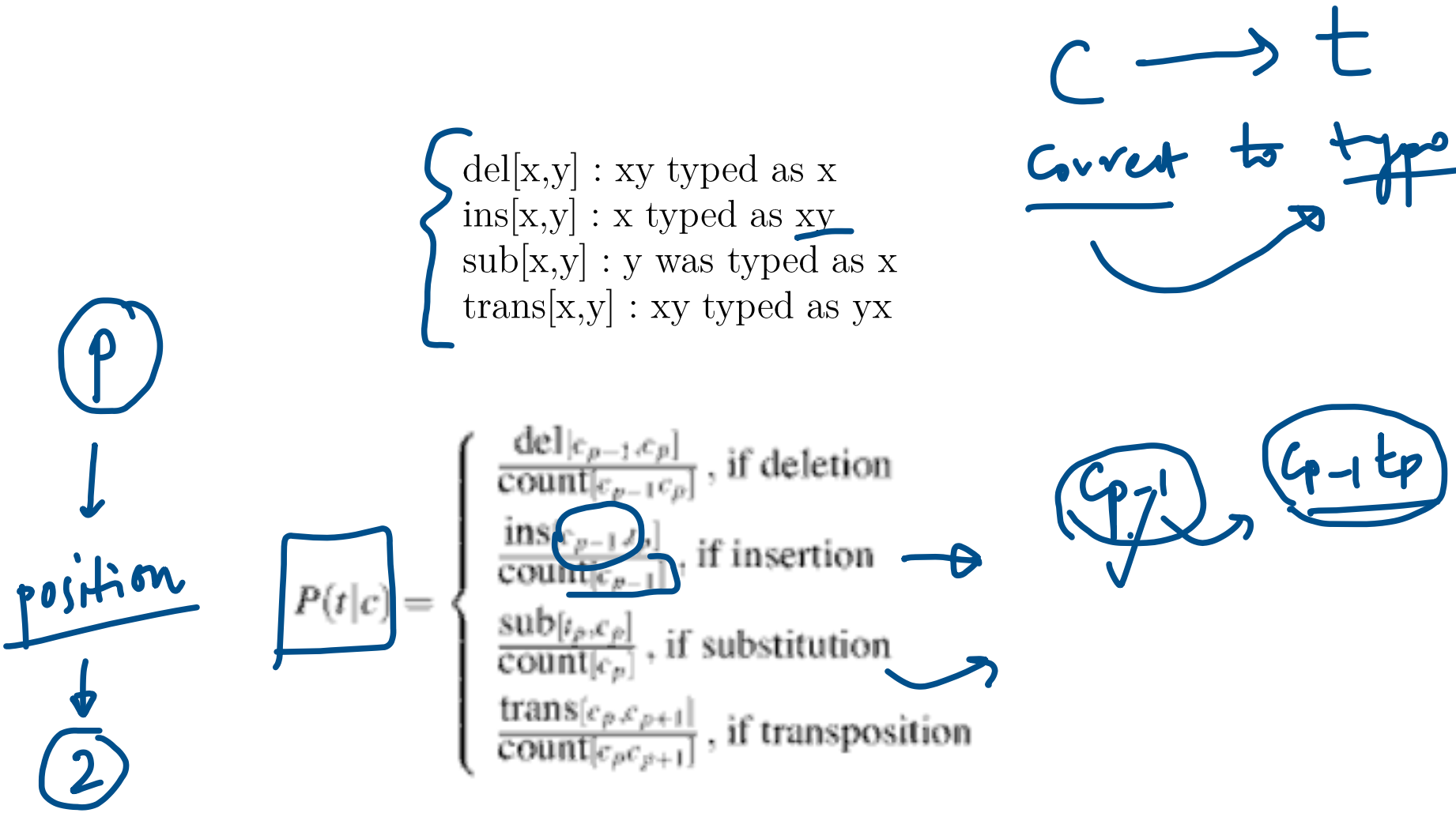
c: apple

t: apple

$t_p \rightarrow$ 'b'

$c_{p-1} \rightarrow$ 'p'

Estimating p(t|c)



sub[X, Y] = Substitution of X (incorrect) for Y (correct)

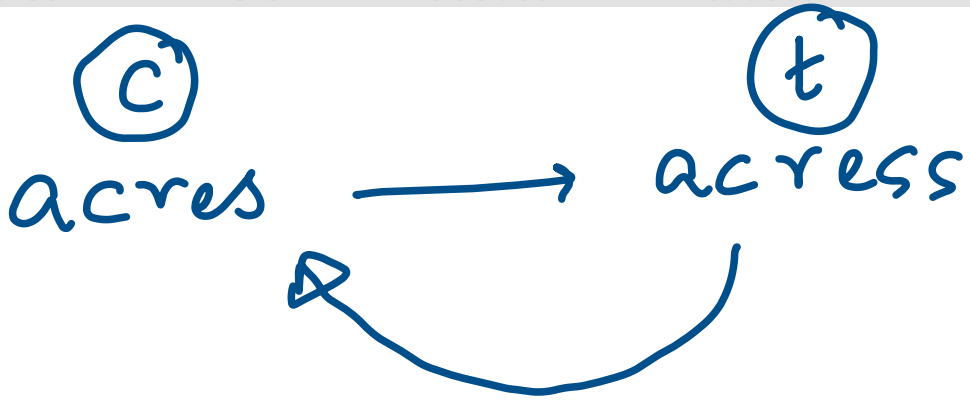
X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Ack: Golding paper, 95

Evaluating Candidates

c	freq(c)	p(c)	p(t|c)	p(t|c)p(c)	%
actress	1343	.0000315	.000117	3.69×10^{-9}	**37%**
cress	0	.000000014	.00000144	2.02×10^{-14}	**0%**
caress	4	.0000001	.00000164	1.64×10^{-13}	**0%**
access	2280	.000058	.000000209	1.21×10^{-11}	**0%**
across	8436	.00019	.0000093	1.77×10^{-9}	**18%**
acres	2879	.000065	.0000321	2.09×10^{-9}	**21%**
acres	2879	.000065	.0000342	2.22×10^{-9}	**23%**

44%



Knowledge Source for Confusion matrices

- There are lists available on Wikipedia and from Roger Mitton (http://www.dcs.bbk.ac.uk/~ROGER/corpora.html) and Peter Norvig (http://norvig. com/ngrams/)

But was that right?

. . . was called a "stellar and versatile acress whose combination of sass and glamour has defined her. . . ".

Reference paper:

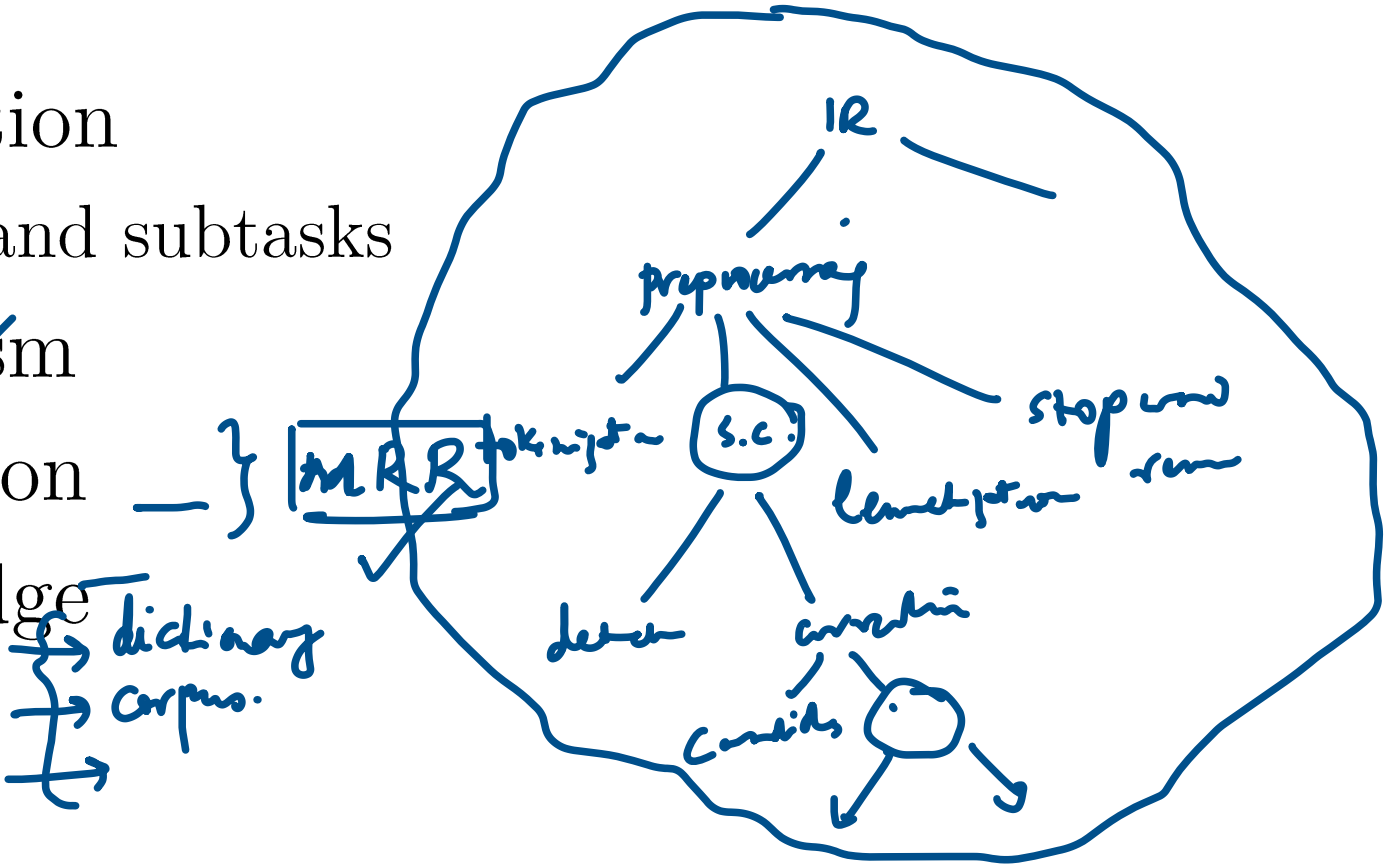
- Kernighan, M. D., Church, K. W. and Gale, W. A. (1990), "***A Spelling Correction Program Based on*** a Noisy Channel Model", Proceedings of COLING '90, Helsinki

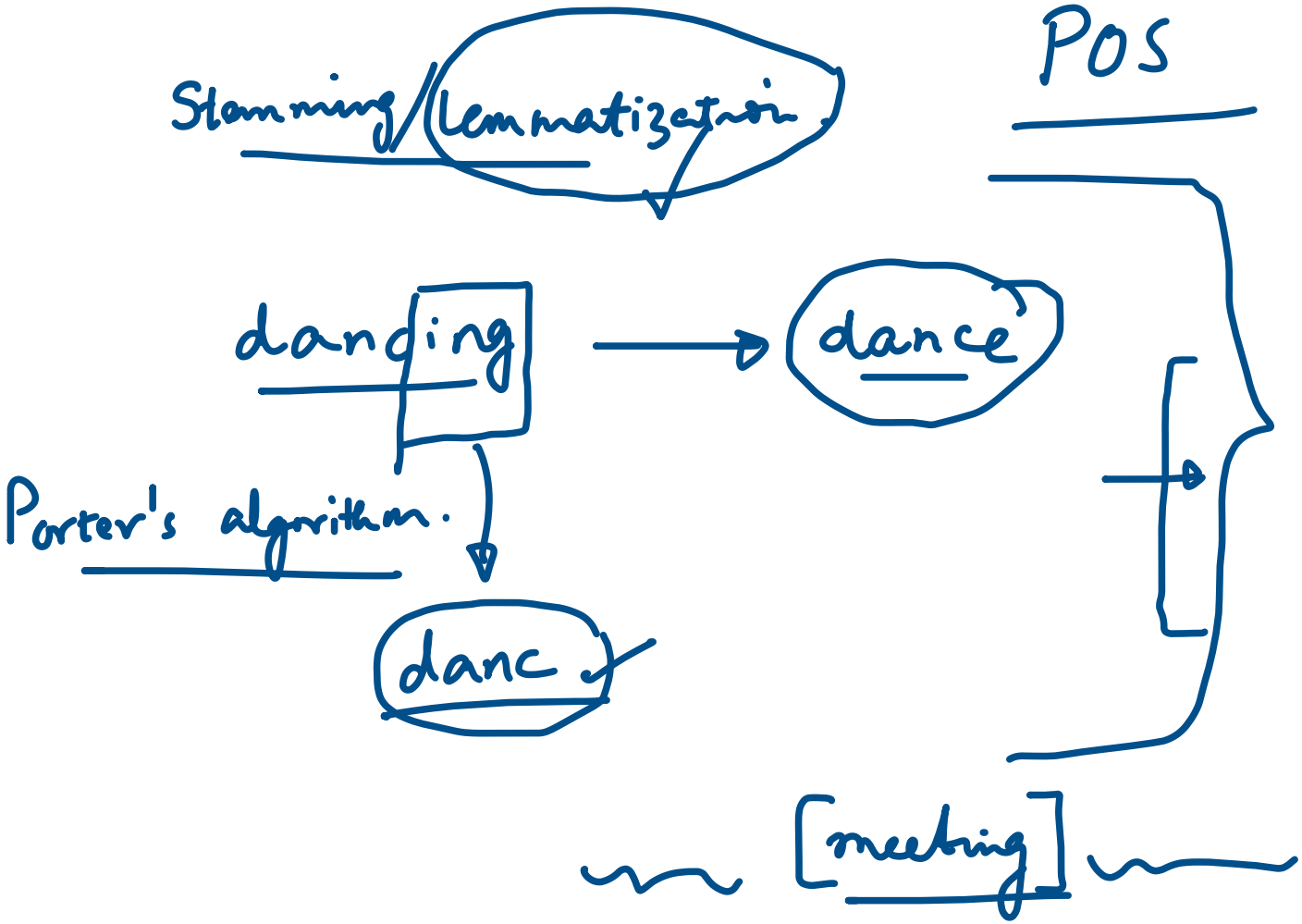
Building a spellcheck application

- Domain
- Application
  - Tasks and subtasks
- Formalism
- Evaluation
- Knowledge

Building an NLP application

- Domain
- Application
  - Tasks and subtasks
- Formalism
- Evaluation
- Knowledge




Stemming / Lemmatization
POS
dancing
dance
Porter's algorithm.
danc
[meeting]