CS344: Introduction to Artificial Intelligence

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 24-25: Argmax Based Computation

Two Views of NLP and the Associated Challenges

1. Classical View
2. Statistical/Machine Learning View

Example of Sentence labeling: Parsing

$[_{S1}[_{S}[_{S}[_{VP}[_{VB}$Come$][_{NP}[_{NNP}$July$]]]]]$
$[_{,}$,$]$
$[_{CC}$ and$]$
$[_{S}$ $[_{NP}$ $[_{DT}$ the$]$ $[_{JJ}$ IIT$]$ $[_{NN}$ campus$]]$
$[_{VP}$ $[_{AUX}$ is$]$
$[_{ADJP}$ $[_{JJ}$ abuzz$]$
$[_{PP}[_{IN}$ with$]$
$[_{NP}[_{ADJP}$ $[_{JJ}$ new$]$ $[_{CC}$ and$]$ $[$ $_{VBG}$ returning$]]$
$[_{NNS}$ students$]]]]]]$
$[_{.}$.$]]]$

Modeling Through the Noisy Channel: Argmax based computation

5 problems in NLP

Foundation: the Noisy Channel Model

The problem formulation is based on the Noisy Channel Model

$(w_n, w_{n-1}, \ldots, w_1)$ $(t_m, t_{m-1}, \ldots, t_1)$

Sequence *w* is transformed into sequence *t*.

$$w^* | t^* = \arg\max_{w} P(w|t)$$

Guess at the correct sequence

Correct sequence

Noisy transformation

Bayesian Decision Theory

- Bayes Theorem : Given the random variables A and B,

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

$P(A \mid B)$ Posterior probability

$P(A)$ Prior probability

$P(B \mid A)$ Likelihood

Bayes Theorem Derivation

$$P(A \cap B) = P(B \cap A)$$

Commutativity of “intersection”

$$P(A)\ P(B \mid A) = P(B)\ P(A \mid B)$$

$$\Rightarrow P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

To understand when and why to apply Bayes Theorem

An example: *it is known that in a population, 1 in 50000 has meningitis and 1 in 20 has stiff neck. It is also observed that 50% of the meningitis patients have stiff neck.*

A doctor observes that a patient has stiff neck. What is the probability that the patient has meningitis?

(Mitchel, Machine Learning, 1997)

Ans: We need to find

P(m|s): probability of meningitis given the stiff neck

Apply Bayes Rule (why?)

$$P(m|s) = [P(m).\ P(s|m)]/P(s)$$

P(m)= prior probability of meningitis

P(s|m)=likelihod of stiff neck given meningitis

P(s)=Probability of stiff neck

Probabilities

$P(m) = \frac{1}{50000}$ ← Prior

$P(s) = \frac{1}{20}$

$P(s|m) = 0.5$ ← Likelihood

$$P(m \mid s) = \frac{P(m)P(s \mid m)}{P(s)} = \frac{\frac{1}{50000} * 0.5}{\frac{1}{20}} = \frac{1}{5000}$$

posterior

$P(m \mid s) \ll P(\sim m \mid s)$ Hence meningitis is not likely

Some Issues

- *p(m/s)* could have been found as

$$\frac{\#(m \cap s)}{\#s}$$

Questions:

- Which is more reliable to compute, *p(s/m)* or *p(m/s)*?
- Which evidence is more sparse , *p(s/m)* or *p(m/s)*?
- Test of significance : The counts are always on a sample of population. Which probability count has sufficient statistics?

5 problems in NLP whose probabilistic formulation use Bayes theorem

The problems

- Part of Speech Tagging: *discussed in detail in subsequent classes*
- Statistical Spell Checking
- Automatic Speech Recognition
- Probabilistic Parsing
- Statistical Machine Translation

Some general observations

$$A^* = \underset{A}{\operatorname{argmax}}\,[P(A|B)]$$

$$= \underset{A}{\operatorname{argmax}}\,[P(A).P(B|A)]$$

Computing and using P(A) and P(B|A), both need

(i) *looking at the **internal structures** of A and B*

(ii) *making independence assumptions*

(iii) *putting together a computation from smaller parts*

PoS tagging: Example

Sentence:

The national committee remarked on a number of other issues.

Tagged output:

The/DET national/ADJ committee/NOU remarked/VRB *on*/PRP *a*/DET *number*/NOU *of*/PRP *other*/ADJ *issues*/NOU*.*

POS Tagging

Best tag *t,**

$$t^* = \arg\max_t P(t \mid w)$$

$$t^* = \prod_1^{N+1} P(t_i \mid t_{i-1}, t_{i-2}) P(w_i \mid t_i)$$

Spell checker: apply Bayes Rule

$$W^* = \text{argmax}\,[P(W|T)]$$
$$= \text{argmax}\,[P(W).P(T|W)]$$

W=correct word, T=misspelt word

- Why apply Bayes rule?
  - Finding $p(w/t)$ vs. $p(t/w)$?
- Assumptions :
  - *t* is obtained from *w* by a single error.
  - The words consist of only alphabets

(Jurafsky and Martin, Speech and NLP, 2000)

4 Confusion Matrices: *sub, ins, del* and *trans*

- If x and y are alphabets,
  - sub(x,y) = # times y is written for x (substitution)
  - ins(x,y) = # times x is written as xy
  - del(x,y) = # times xy is written as x
  - trans(x,y) = # times xy is written as yx

Probabilities from confusion matrix

- $P(t/w) = P(t/w)_S + P(t/w)_I + P(t/w)_D + P(t/w)_X$

where

$P(t|w)_S$ = sub(x,y) / count of x

$P(t|w)_I$ = ins(x,y) / count of x

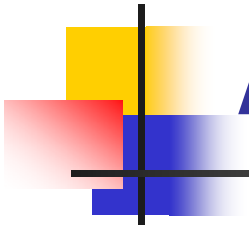
$P(t|w)_D$ = del(x,y) / count of x

$P(t|w)_X$ = trans(x,y) / count of x

- These are considered to be mutually exclusive events

URLs for database of misspelt words

- http://www.wsu.edu/~brians/errors/errors.html
- http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines



A sample

- abandonned->abandoned
- aberation->aberration
- abilties->abilities
- abilty->ability
- abondon->abandon
- abondoned->abandoned
- abondoning->abandoning
- abondons->abandons
- aborigene->aborigine

fi yuo cna raed tihs, yuo hvae a sgtrane mnid too.
Cna yuo raed tihs? Olny 55 plepoe can.

i cdnuolt blveiee taht I cluod aulaclty uesdnatnrd waht I was rdanieg.
The phaonmneal pweor of the hmuan mnid, aoccdrnig to a rscheearch at
Cmabrigde Uinervtisy, it dseno't mtaetr in waht oerdr the ltteres in a wrod are, the olny iproamtnt tihng is taht the frsit and lsat ltteer be in the rghit pclae. The rset can be a taotl mses and you can sitll raed
it whotuit a pboerlm. Tihs is bcuseae the huamn mnid deos not raed
ervey lteter by istlef, but the wrod as a wlohe. Azanmig huh? yaeh and
I

awlyas tghuhot slpeling was ipmorantt! if you can raed tihs forwrad it.

Spell checking: Example

- Given *aple*, find and rank
  - P(maple|aple), P(apple|aple), P(able|aple), P(pale|aple) etc.

- *Exercise*: *Give an intuitive feel for which of these will rank higher*

Problem 3: Probabilistic Speech Recognition

- **Problem Definition : Given a sequence of speech signals, identify the words.**
- **2 steps :**
  - **Segmentation (Word Boundary Detection)**
  - **Identify the word**
- **Isolated Word Recognition :**
  - **Identify W given SS (speech signal)**

$$\hat{W} = \arg\max_{W} P(W \mid SS)$$

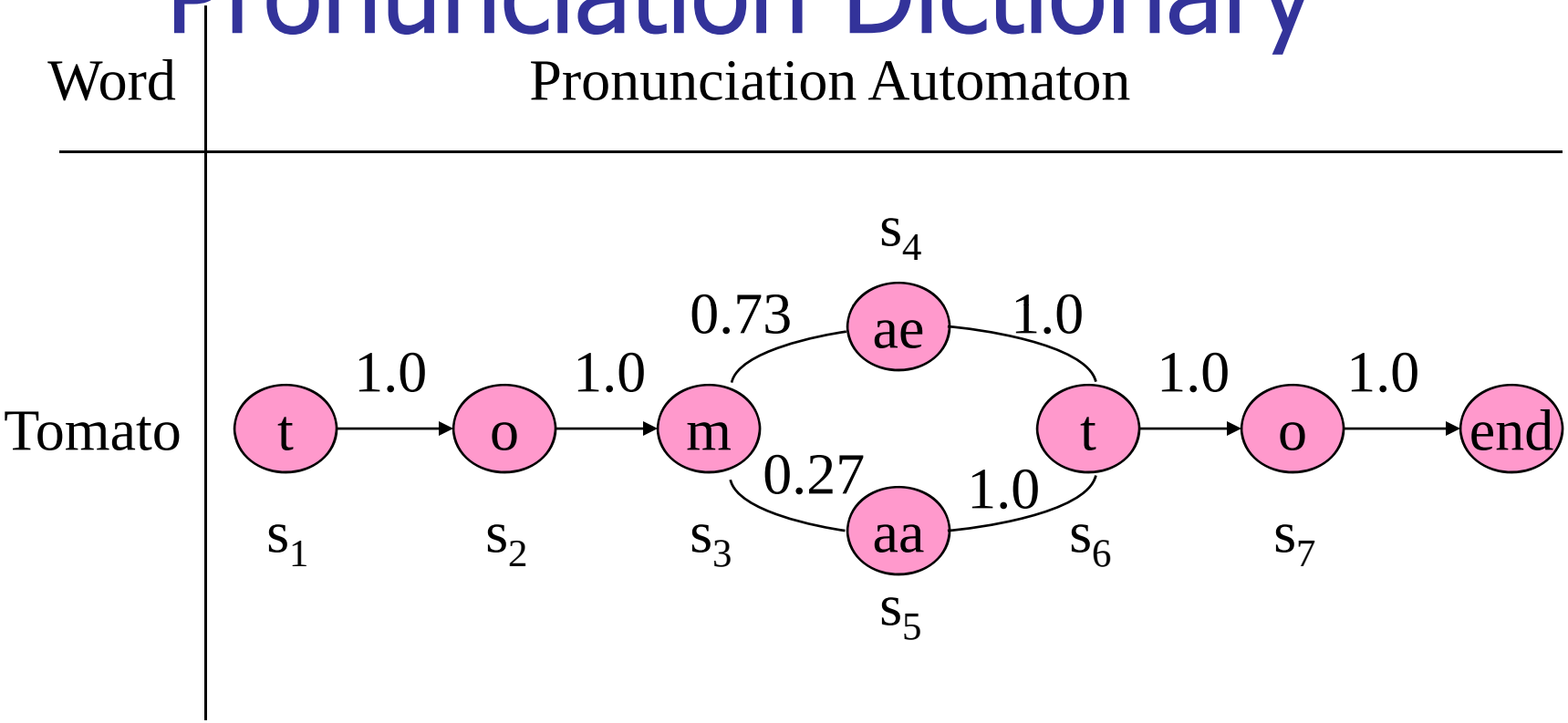
Speech recognition: Identifying the word

$$\hat{W} = \arg\max_{W} P(W \mid SS)$$
$$= \arg\max_{W} P(W)P(SS \mid W)$$

- *P(SS/W)* = likelihood called "phonological model " → intuitively more tractable!
- *P(W)* = prior probability called "language model"

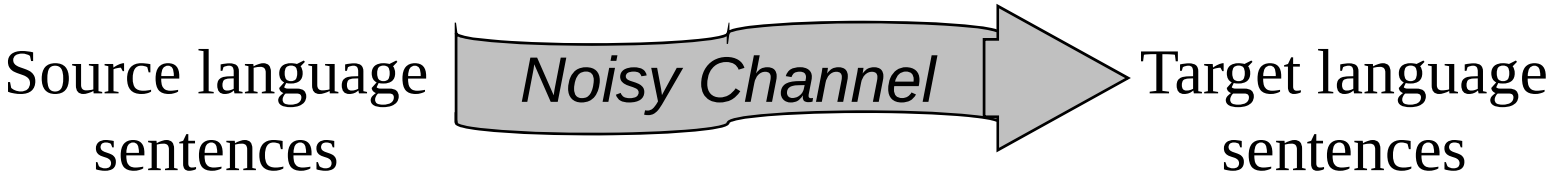
$$P(W) = \frac{\#\ \text{W appears in the corpus}}{\#\ \text{words in the corpus}}$$

Pronunciation Dictionary



- *P(SS/W)* is maintained in this way.
- *P(t o m ae t o /Word is "tomato")* = Product of arc probabilities

Problem 4: Statistical Machine Translation



- What sentence in the target language will maximise the probability

  P(target sentence|source sentence)

Statistical MT: Parallel Texts

- Parallel texts
  - Instruction manuals
  - Hong Kong legislation
  - Macao legislation
  - Canadian parliament Hansards
  - United nation reports
  - Official journal of the European Communities
  - Trilingual documents in Indian states

SMT: formalism

- Source language: F
- Target language: E
- Source language sentence: f
- Target language sentence: e
- Source language word: w^f
- Target language word: w^e

SMT Model

- To translate *f*:
  - Assume that all sentences in *E* are translations of *f* with some probability!
  - Choose the translation with the highest probability

$$\hat{e} = \arg\max_{e} \; (p(e \mid f))$$

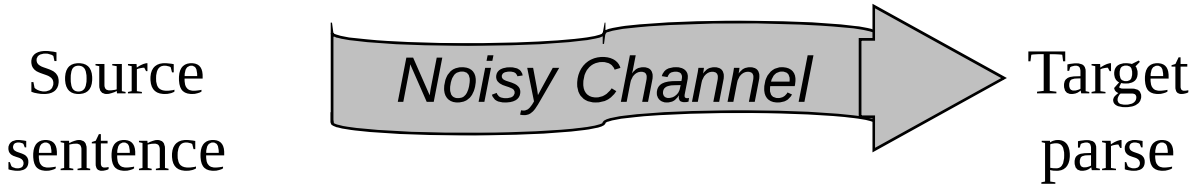
SMT: Apply Bayes Rule

$$\hat{e} = \arg\max_{e} \; (p(e).p(f \mid e))$$

P(e) is called the ***language model*** and
stands for ***fluency***
and
P(f|e} is called the ***translation model*** and
stands for ***faithfulness***

Both these are computed by breaking them down
Into smaller components of n-grams

Problem 5: Parsing



$$T^* = \underset{T}{\operatorname{argmax}}\,[P(T|S)]$$

$$= \underset{T}{\operatorname{argmax}}\,[P(T).P(S|T)]$$

$$= \underset{T}{\operatorname{argmax}}\,[P(T)]$$

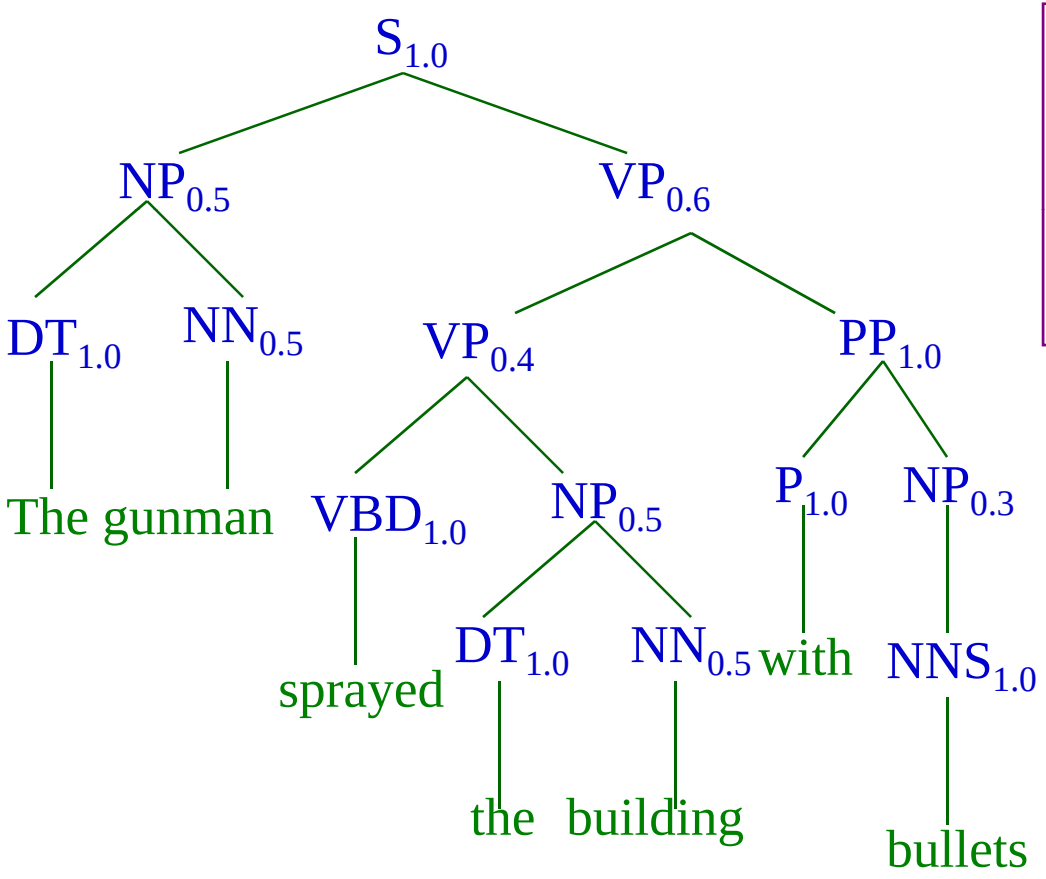
since given the parse the sentence is completely determined and $P(S|T)=1$

Probabilistic Context Free Grammars

- S $\rightarrow$ NP VP 1.0
- NP $\rightarrow$ DT NN 0.5
- NP $\rightarrow$ NNS 0.3
- NP $\rightarrow$ NP PP 0.2
- PP $\rightarrow$ P NP 1.0
- VP $\rightarrow$ VP PP 0.6
- VP $\rightarrow$ VBD NP 0.4
- DT $\rightarrow$ the 1.0
- NN $\rightarrow$ gunman 0.5
- NN $\rightarrow$ building 0.5
- VBD $\rightarrow$ sprayed 1.0
- NNS $\rightarrow$ bullets 1.0

Example Parse t_1`

- The gunman sprayed the building with bullets.



$P(t_1) = 1.0 * 0.5 * 1.0 * 0.5 * 0.6 * 0.4 * 1.0 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0 =$
0.00225

Is NLP Really Needed

Post-1

- POST----5 TITLE: "Wants to invest in IPO? Think again" |

Hereâ€™s a sobering thought for those who believe in investing in IPOs. Listing gains â€” the return on the IPO scrip at the close of listing day over the allotment price â€” have been falling substantially in the past two years. Average listing gains have fallen from 38% in 2005 to as low as 2% in the first half of 2007.Of the 159 book-built initial public offerings (IPOs) in India between 2000 and 2007, two-thirds saw listing gains. However, these gains have eroded sharply in recent years.Experts say this trend can be attributed to the aggressive pricing strategy that investment bankers adopt before an IPO. â€œWhile the drop in average listing gains is not a good sign, it could be due to the fact that IPO issue managers are getting aggressive with pricing of the issues,â€ says Anand Rathi, chief economist, Sujan Hajra.While the listing gain was 38% in 2005 over 34 issues, it fell to 30% in 2006 over 61 issues and to 2% in 2007 till mid-April over 34 issues. The overall listing gain for 159 issues listed since 2000 has been 23%, according to an analysis by Anand Rathi Securities.Aggressive pricing means the scrip has often been priced at the high end of the pricing range, which would restrict the upward movement of the stock, leading to reduced listing gains for the investor. It also tends to suggest investors should not indiscriminately pump in money into IPOs.But some market experts point out that India fares better than other countries. â€œInternationally, there have been periods of negative returns and low positive returns in India should not be considered a bad thing.

Post-2

- POST----7TITLE: "[IIM-Jobs] ***** Bank: International Projects Group - Manager"|
Please send your CV & cover letter to anup.abraham@*****bank.com ***** Bank, through its International Banking Group (IBG), is expanding beyond the Indian market with an intent to become a significant player in the global marketplace. The exciting growth in the overseas markets is driven not only by India linked opportunities, but also by opportunities of impact that we see as a local player in these overseas markets and / or as a bank with global footprint. IBG comprises of Retail banking, Corporate banking & Treasury in 17 overseas markets we are present in. Technology is seen as key part of the business strategy, and critical to business innovation & capability scale up. The International Projects Group in IBG takes ownership of defining & delivering business critical IT projects, and directly impact business growth. Role: Manager Â– International Projects Group Purpose of the role: Define IT initiatives and manage IT projects to achieve business goals. The project domain will be retail, corporate & treasury. The incumbent will work with teams across functions (including internal technology teams & IT vendors for development/implementation) and locations to deliver significant & measurable impact to the business. Location: Mumbai (Short travel to overseas locations may be needed) Key Deliverables: Conceptualize IT initiatives, define business requirements

Sentiment Classification

- Positive, negative, neutral – 3 class
- Sports, economics, literature - multi class
- Create a representation for the document
- Classify the representation

The most popular way of representing a document is feature vector (indicator sequence).

Established Techniques

- Naïve Bayes Classifier (NBC)
- Support Vector Machines (SVM)
- Neural Networks
- K nearest neighbor classifier
- Latent Semantic Indexing
- Decision Tree ID3
- Concept based indexing

Successful Approaches

The following are successful approaches as reported in literature.

- NBC – simple to understand and implement
- SVM – complex, requires foundations of perceptions

Mathematical Setting

We have training set

A: Positive Sentiment Docs

B: Negative Sentiment Docs

Indicator/feature vectors to be formed

Let the class of positive and negative documents be C_+ and C_- , respectively.

Given a new document **D** label it positive if

$$P(C_+|D) > P(C_-|D)$$

Priori Probability

Document	Vector	Classification
D1	V1	+
D2	V2	-
D3	V3	+
..	..	..
D_{4000}	V_{4000}	-

Let T = Total no of documents
And let |+| = M
So,|-| = T-M

P(D being positive)=M/T

Priori probability is calculated without considering any features of the new document.

Apply Bayes Theorem

Steps followed for the NBC algorithm:

- Calculate Prior Probability of the classes. $P(C_+)$ and $P(C_-)$
- Calculate feature probabilities of new document. $P(D|C_+)$ and $P(D|C_-)$
- Probability of a document **D** belonging to a class **C** can be calculated by Baye's Theorem as follows:

$$P(C|D) = \frac{P(C) * P(D|C)}{P(D)}$$

- Document belongs to C_+ , if

$$P(C_+) * P(D|C_+) > P(C_-) * P(D|C_-)$$

Calculating $P(D|C_+)$

- Identify a set of features/indicators to represent a document and generate a feature vector (V_D). $V_D = \langle x_1 , x_2 , x_3 \ldots x_n \rangle$
- Hence, $P(D|C_+) = P(V_D|C_+)$

$$= P(\langle x_1 , x_2 , x_3 \ldots x_n \rangle \mid C_+)$$

$$= \frac{|\langle x1,x2,x3\ldots..xn\rangle, C_+|}{|C_+|}$$

- Based on the assumption that all features are Independently Identically Distributed (IID)

$$= P(\langle x_1 , x_2 , x_3 \ldots x_n \rangle \mid C_+)$$

$$= P(x_1 |C_+) * P(x_2 |C_+) * P(x_3 |C_+) *\ldots. P(x_n |C_+)$$

$$= \prod_{i=1}^{n} P(x_i |C_+)$$

Baseline Accuracy

- Just on Tokens as features, **80%** accuracy
- 20% probability of a document being misclassified
- On large sets this is significant

To improve accuracy…

- Clean corpora
- POS tag
- Concentrate on critical POS tags (e.g. *adjective*)
- Remove ‘objective’ sentences ('of' ones)
- Do aggregation

Use minimal to sophisticated NLP