

# Natural Language Models and Interfaces

BSc Artificial Intelligence

Lecturer: Wilker Aziz

Institute for Logic, Language, and Computation

2018, week 2, lecture a

# NLMI

Probability of a sentence

Language models

Smoothing

Evaluating language models

# Chomsky once said

*It must be recognised that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term* — Chomsky, 1969

- ▶ Chomsky is the father of modern linguistics
- ▶ he also made significant contributions to formal language theory

Yet here we are discussing how we will be assigning probabilities to sentences

- ▶ should we listen to authority?

# Objective probability

Perhaps Chomsky only acknowledges **objective probability**

- ▶ a notion of *frequency* or *propensity*
- ▶ a tendency of a given situation to yield a certain outcome

Example

*The winged Irritator Challengeri pounced upon the hapless Bambiraptor*

- ▶ how many times have you heard this sentence in your lifetime?

Under such a view, Chomsky's observation (or claim) is pretty reasonable

# Subjective probability

Under the **subjective view**, probability has nothing to do with the frequency or propensity of outcomes

- ▶ it is a notion of **reasonable expectation**
- ▶ represents a **state of knowledge**  
or a **quantification of personal belief**

## Example

*The winged Irritator Challengeri pounced upon the hapless Bambiraptor*

- ▶ according to your knowledge of English, is it at all reasonable to think of this as a sentence and expect it may be uttered?

# Subjective vs objective probability

It's unlikely we will ever hear either

- ▶ The winged Irritator Challengeri pounced upon the hapless Bambiraptor
- ▶ Upon the pounced winged hapless Bambiraptor the Challengeri Irritator winged

yet it's far more reasonable to **expect** one than the other

- ▶ and that's why we will not listen to authority at least this time ;)

---

Chomsky's theories and contributions are pretty important though!

# Random sentences

We will model probability distributions over **sentences**!

- ▶ let's start with a vocabulary of **words**  $\Sigma$
- ▶ and let's assume a *language is a subset of strings* in  $\Sigma^*$

# Random sentences

We will model probability distributions over **sentences**!

- ▶ let's start with a vocabulary of **words**  $\Sigma$
- ▶ and let's assume a *language is a subset of strings* in  $\Sigma^*$

Let's start with **random words**

- ▶ define an rv  $X$  that maps from  $\Sigma$  to  $\mathbb{R}$
- ▶ the mapping is an arbitrary enumeration of  $\Sigma$   
*some word is mapped to 1, some other word is mapped to 2, ..., some other word is mapped to  $v = |\Sigma|$*
- ▶ with  $v = |\Sigma|$  we say  
let  $X$  take on values in an index set  $\mathcal{X} = \{1, \dots, v\}$  of  $\Sigma$

▶ Quiz



# Random sentences

We will model probability distributions over **sentences**!

- ▶ let's start with a vocabulary of **words**  $\Sigma$
- ▶ and let's assume a *language is a subset of strings* in  $\Sigma^*$

Let's start with **random words**

- ▶ define an rv  $X$  that maps from  $\Sigma$  to  $\mathbb{R}$
- ▶ the mapping is an arbitrary enumeration of  $\Sigma$   
*some word is mapped to 1, some other word is mapped to 2, ..., some other word is mapped to  $v = |\Sigma|$*
- ▶ with  $v = |\Sigma|$  we say ▶ Quiz  
let  $X$  take on values in an index set  $\mathcal{X} = \{1, \dots, v\}$  of  $\Sigma$

A random sentence  $S$  of length  $n$  is a **sequence** of random words  
 $\langle X_1, \dots, X_n \rangle$  which we also denote  $X_1^n$

# Probability of a sentence

The probability of a sentence  $\langle x_1, \dots, x_n \rangle$

$$P_S(\langle x_1, \dots, x_n \rangle) = P_N(n) \underbrace{\prod_{i=1}^n P_{X|H}(x_i | x_{<i})}_{\text{chain rule}}$$

- ▶  $x_1^n$  shorthand for the **sequence**  $\langle x_1, \dots, x_n \rangle$   
and  $x_1^n$  is  $\langle x_1 \rangle$  if  $n = 1$
- ▶  $x_{<i}$  shorthand for the **prefix sequence**  $\langle x_1, \dots, x_{i-1} \rangle$   
and  $x_{<i}$  is the empty sequence  $\langle \rangle$  if  $i = 1$
- ▶ we call the random sequence  $H$  the **history**  
like  $S$ , its sample space is  $\Sigma^*$

# Probability of a sentence

The probability of a sentence  $\langle x_1, \dots, x_n \rangle$

$$P_S(\langle x_1, \dots, x_n \rangle) = P_N(n) \underbrace{\prod_{i=1}^n P_{X|H}(x_i | x_{<i})}_{\text{chain rule}}$$

- ▶  $x_1^n$  shorthand for the **sequence**  $\langle x_1, \dots, x_n \rangle$   
and  $x_1^n$  is  $\langle x_1 \rangle$  if  $n = 1$
- ▶  $x_{<i}$  shorthand for the **prefix sequence**  $\langle x_1, \dots, x_{i-1} \rangle$   
and  $x_{<i}$  is the empty sequence  $\langle \rangle$  if  $i = 1$
- ▶ we call the random sequence  $H$  the **history**  
like  $S$ , its sample space is  $\Sigma^*$

A sentence is a **structured object** — for example it has an *order*

- ▶ above is a **factorisation** of the distribution  $P_S(S)$
- ▶ it *chops* the structure **generating** one piece (a word) at a time
- ▶ a factor  $P_{X|H}$  is a **conditional probability distribution** (cpd)

# Generative story

The stochastic procedure that yields a sentence is

▶ also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

▶  $X_i | x_{<i} \sim P_{X|H}$

# Generative story

The stochastic procedure that yields a sentence is

► also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

►  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

►  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

# Generative story

The stochastic procedure that yields a sentence is

► also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

►  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

►  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

# Generative story

The stochastic procedure that yields a sentence is

► also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

►  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

►  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

►  $\langle \text{this}, X_2, X_3, X_4, X_5 \rangle$

# Generative story

The stochastic procedure that yields a sentence is

► also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

►  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

►  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

►  $\langle \text{this}, X_2, X_3, X_4, X_5 \rangle$

►  $\langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$



# Generative story

The stochastic procedure that yields a sentence is

- ▶ also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

- ▶  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

- ▶  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

- ▶  $\langle \text{this}, X_2, X_3, X_4, X_5 \rangle$

- ▶  $\langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$

- ▶  $\langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$

# Generative story

The stochastic procedure that yields a sentence is

- ▶ also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

- ▶  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

- ▶  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

- ▶  $\langle \text{this}, X_2, X_3, X_4, X_5 \rangle$

- ▶  $\langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$

- ▶  $\langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$

- ▶  $\langle \text{this}, \text{is}, \text{a}, \text{short}, X_5 \rangle$

# Generative story

The stochastic procedure that yields a sentence is

► also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

►  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

►  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

►  $\langle \text{this}, X_2, X_3, X_4, X_5 \rangle$

►  $\langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$

►  $\langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$

►  $\langle \text{this}, \text{is}, \text{a}, \text{short}, X_5 \rangle$

►  $\langle \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence} \rangle$

# Generative story

The stochastic procedure that yields a sentence is

▶ also known as **generative story**

1. Sample a length  $N \sim P_N$

2. For  $i = 1, \dots, n$

▶  $X_i | x_{<i} \sim P_{X|H}$

Example:  $N \sim P_N$

▶  $\langle X_1, X_2, X_3, X_4, X_5 \rangle$

$X_i | x_{<i} \sim P_{X|H}$  for  $i = 1, \dots, 5$

▶  $\langle \text{this}, X_2, X_3, X_4, X_5 \rangle$

▶  $\langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$

▶  $\langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$

▶  $\langle \text{this}, \text{is}, \text{a}, \text{short}, X_5 \rangle$

▶  $\langle \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence} \rangle$

Quiz

# Notation mess

You will often find in LM literature

$$\underbrace{P(x_1, \dots, x_n)}_{\text{joint probability}} = \underbrace{P(x_1) \prod_{i=2}^n P(x_i | x_1, \dots, x_{i-1})}_{\text{chain rule}}$$

- ▶ joint distribution does not care about the order of its arguments
  - ▶ thus this actually hides that  $x_1, \dots, x_n$  is a **sequence**
  - ▶ indices are naming the rvs (not an arbitrary enumeration)

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, X_3 = x_3) &= P(X_3 = x_3, X_1 = x_1, X_2 = x_2) \\ &= P_{X_1}(x_1) P_{X_2|X_1}(x_2|x_1) P_{X_3|X_1 X_2}(x_3|x_1, x_2) \end{aligned}$$

- ▶ while a correct application of chain rule, in a modelling context, this hides the fact that the **length** of the sequence is itself random

# Less ambiguous notation

Instead of

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i | x_1, \dots, x_{i-1})$$

We prefer

- ▶  $\langle x_1, \dots, x_n \rangle$  or  $x_1^n$  for sequences
- ▶  $\langle x_1, \dots, x_{i-1} \rangle$  or  $x_{<i}$  for prefix sequences

and the joint probability factorises as

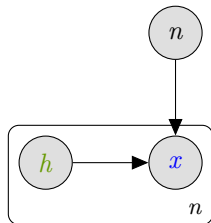
$$P_S(x_1^n) = P_N(n) P_{S|N}(x_1^n | n)$$

- ▶ with  $P_{S|N}(x_1^n | n) = \prod_{i=1}^n P_{X|H}(x_i | x_{<i})$

# Language models are directed graphical models

Probabilistic directed graphical model (or Bayesian net)

- ▶ directed acyclic graph
- ▶ nodes are **random variables**
- ▶ a plate represents a loop
- ▶ a directed arrow denotes **conditional dependence**



$$P_S(\langle x_1, \dots, x_n \rangle) = P_N(n) \prod_{i=1}^n P_{X|H}(x_i | x_{<i})$$

# NLMI

Probability of a sentence

Language models

Smoothing

Evaluating language models



# Parameterisation

We are very close to defining a complete model

- ▶ we need to choose parametric families for  $P_N$  and  $P_{X|H}$

Length distribution  $P_N$

- ▶ for simplicity we pick some uniform distribution  
thus  $P_N(n) = c$

Next word distribution  $P_{X|H}$

- ▶ we start by making a **very unrealistic** simplifying assumption
  - ▶ we assume the next word is independent of the history  $X \perp H$   
i.e.  $P_{X|H} = P_X$
- ▶ and make  $P_X$  a categorical distribution  $\text{Cat}(\theta_1, \dots, \theta_v)$

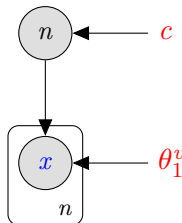
---

Realistic length distributions can be rather complex (Sichel 1974)

# Unigram language model

## Unigram factorisation and categorical parameterisation

$$\begin{aligned}P_S(\langle x_1, \dots, x_n \rangle) &= P_N(n) \prod_{i=1}^n P_{X|H}(x_i | x_{<i}) \\&\approx c \prod_{i=1}^n P_X(x_i) \\&\propto \prod_{i=1}^n \text{Cat}(X = x_i | \theta_1, \dots, \theta_v)\end{aligned}$$



Recall the Categorical pmf

$$\blacktriangleright \text{Cat}(X = a | \theta_1, \dots, \theta_v) = \prod_{x=1}^v \theta_x^{\delta_{xa}}$$

# Unigram LM - illustration

$$N \sim P_N$$

# Unigram LM - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

# Unigram LM - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

$$X_i | x_{<i} \approx \langle \rangle \sim P_X \text{ for } i = 1, \dots, 5$$

# Unigram LM - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

$$X_i | x_{<i} \approx \langle \rangle \sim P_X \text{ for } i = 1, \dots, 5$$

$$\blacktriangleright \langle \text{this}, X_2, X_3, X_4, X_5 \rangle$$

# Unigram LM - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

$$X_i | x_{<i} \approx \langle \rangle \sim P_X \text{ for } i = 1, \dots, 5$$

$$\blacktriangleright \langle \text{this}, X_2, X_3, X_4, X_5 \rangle$$

$$\blacktriangleright \langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$$

# Unigram LM - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

$$X_i | x_{<i} \approx \langle \rangle \sim P_X \text{ for } i = 1, \dots, 5$$

$$\blacktriangleright \langle \text{this}, X_2, X_3, X_4, X_5 \rangle$$

$$\blacktriangleright \langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$$

$$\blacktriangleright \langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$$



# Unigram LM - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

$$X_i | x_{<i} \approx \langle \rangle \sim P_X \text{ for } i = 1, \dots, 5$$

$$\blacktriangleright \langle \text{this}, X_2, X_3, X_4, X_5 \rangle$$

$$\blacktriangleright \langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$$

$$\blacktriangleright \langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$$

$$\blacktriangleright \langle \text{this}, \text{is}, \text{a}, \text{short}, X_5 \rangle$$

# Unigram LM - illustration

$$N \sim P_N$$

$$\triangleright \langle X_1, X_2, X_3, X_4, X_5 \rangle$$

$$X_i | x_{<i} \approx \langle \rangle \sim P_X \text{ for } i = 1, \dots, 5$$

$$\triangleright \langle \text{this}, X_2, X_3, X_4, X_5 \rangle$$

$$\triangleright \langle \text{this}, \text{is}, X_3, X_4, X_5 \rangle$$

$$\triangleright \langle \text{this}, \text{is}, \text{a}, X_4, X_5 \rangle$$

$$\triangleright \langle \text{this}, \text{is}, \text{a}, \text{short}, X_5 \rangle$$

$$\triangleright \langle \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence} \rangle$$

# Let's see what's wrong with the unigram model

Consider the probability of the sentences

- ▶ the winged irritator challenger<sub>i</sub> pounced upon the hapless bambiraptor
- ▶ upon the pounced hapless bambiraptor the challenger<sub>i</sub> irritator winged

Does the model quantify a *reasonable* expectation?

# Let's see what's wrong with the unigram model

Consider the probability of the sentences

- ▶ the winged irritator challenger<sub>i</sub> pounced upon the hapless bambiraptor
- ▶ upon the pounced hapless bambiraptor the challenger<sub>i</sub> irritator winged

Does the model quantify a *reasonable* expectation? **No!**

# Let's see what's wrong with the unigram model

Consider the probability of the sentences

- ▶ the winged irritator challenger<sub>i</sub> pounced upon the hapless bambiraptor
- ▶ upon the pounced hapless bambiraptor the challenger<sub>i</sub> irritator winged

Does the model quantify a *reasonable* expectation? **No!**

- ▶ both sequences have the exact same unigrams
- ▶ and they occur exactly the same number of times

# Let's see what's wrong with the unigram model

Consider the probability of the sentences

- ▶ the winged irritator challenger<sub>i</sub> pounced upon the hapless bambiraptor
- ▶ upon the pounced hapless bambiraptor the challenger<sub>i</sub> irritator winged

Does the model quantify a *reasonable* expectation? No!

- ▶ both sequences have the exact same unigrams
- ▶ and they occur exactly the same number of times

Unigram language models see a sentence as a *mset*

- ▶ but before we fix them, let's get to estimation of  $\theta_1^v$

# MLE for unigram LMs

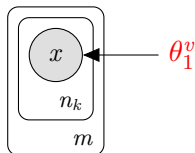
The key cpd in the unigram LM is

$$P_X(X; \theta_1^v) = \text{Cat}(\theta_1, \dots, \theta_v)$$

which is specified by  $v$  **parameters** (**word probabilities**)

Say we have a dataset of  $m$  observations  $\left( \langle x_1^{(k)}, \dots, x_{n_k}^{(k)} \rangle \right)_{k=1}^m$

► what's the MLE solution for  $\theta_x$ ?



# MLE for unigram LMs

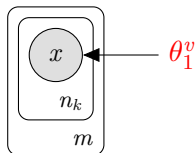
The key cpd in the unigram LM is

$$P_X(X; \theta_1^v) = \text{Cat}(\theta_1, \dots, \theta_v)$$

which is specified by  $v$  **parameters** (**word probabilities**)

Say we have a dataset of  $m$  observations  $\left(\langle x_1^{(k)}, \dots, x_{n_k}^{(k)} \rangle\right)_{k=1}^m$

► what's the MLE solution for  $\theta_x$ ?



$$\begin{aligned}\theta_x &= \frac{\sum_{k=1}^m \sum_{i=1}^{n_k} [x = x_i^{(k)}]}{\sum_{k=1}^m n_k} \\ &= \frac{\text{count}(x)}{\text{number of tokens}}\end{aligned}$$



# How do we fix a unigram LM?

We need to relax our **way-too-strong** independence assumption

- ▶ that is, we need to condition on history

# How do we fix a unigram LM?

We need to relax our **way-too-strong** independence assumption

- ▶ that is, we need to condition on history

but there's a problem with conditioning on the complete history

- ▶ and to understand it we need to look into **tabular cpds**

# Tabular cpds

A tabular cpd is a set of probability distributions,  
for each **conditioning context** we get a new distribution

Recall the example involving grades and recommendation letters

$P_{G L}$				$P_{G L}$			
Letter ( $L$ )	Grade ( $G$ )			Letter ( $L$ )	Grade ( $G$ )		
	[0, 6)	[6, 8)	[8, 10]		[0, 6)	[6, 8)	[8, 10]
	1	2	3		1	2	3
0	0.27	0.71	0.02	0	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$
1	0.10	0.68	0.22	1	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$

Table : Conditional distribution:  $G|L = l \sim \text{Cat}(\theta_{l,1}, \dots, \theta_{l,3})$

# Data sparsity

If we have 1 cpd per assignment of the conditioning variable  
how many cpds do we need to estimate an LM with full history?

# Data sparsity

If we have 1 cpd per assignment of the conditioning variable  
how many cpds do we need to estimate an LM with full history?

- ▶ infinitely many!
- ▶ there are no reasonable limits to what histories can be

# Data sparsity

If we have 1 cpd per assignment of the conditioning variable  
how many cpds do we need to estimate an LM with full history?

- ▶ infinitely many!
- ▶ there are no reasonable limits to what histories can be

The more parameters we need to estimate,  
the more data we need

- ▶ many valid histories will never be seen

*The winged Irritator Challengeri pounced* →?

# Conditional independence

$o$ th order Markov assumption

- ▶ we forget some—but not all—history
- ▶ make **next word** independent of **all but**  $o$  preceding words
- ▶ we call this class of models  $n$ -gram language models  
*we use  $o = n - 1$  to avoid confusion with sentence length*

Next word distribution  $X_i | x_{<i} \approx x_{i-o}^{i-1} \sim \text{Cat}(\theta_1^{(h_i)}, \dots, \theta_v^{(h_i)})$

- ▶ with  $x_{i-o}^{i-1}$  a shorthand for  $\langle x_{i-o}, \dots, x_{i-1} \rangle$   
 $\langle \rangle$  if  $o = 0$  and  $\langle x_{i-1} \rangle$  if  $o = 1$
- ▶  $h_i$  uniquely identifies the  $i$ th shortened random history

# Conditional independence

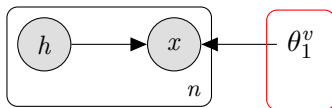
$o$ th order Markov assumption

- ▶ we forget some—but not all—history
- ▶ make **next word** independent of **all but**  $o$  preceding words
- ▶ we call this class of models  $n$ -gram language models  
*we use  $o = n - 1$  to avoid confusion with sentence length*

Next word distribution  $X_i | x_{<i} \approx x_{i-o}^{i-1} \sim \text{Cat}(\theta_1^{(h_i)}, \dots, \theta_v^{(h_i)})$

- ▶ with  $x_{i-o}^{i-1}$  a shorthand for  $\langle x_{i-o}, \dots, x_{i-1} \rangle$   
 $\langle \rangle$  if  $o = 0$  and  $\langle x_{i-1} \rangle$  if  $o = 1$
- ▶  $h_i$  uniquely identifies the  $i$ th shortened random history

How many cpds can we have?





# Conditional independence

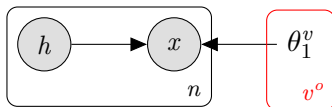
$o$ th order Markov assumption

- ▶ we forget some—but not all—history
- ▶ make **next word** independent of **all but**  $o$  preceding words
- ▶ we call this class of models  $n$ -gram language models  
*we use  $o = n - 1$  to avoid confusion with sentence length*

Next word distribution  $X_i | x_{<i} \approx x_{i-o}^{i-1} \sim \text{Cat}(\theta_1^{(h_i)}, \dots, \theta_v^{(h_i)})$

- ▶ with  $x_{i-o}^{i-1}$  a shorthand for  $\langle x_{i-o}, \dots, x_{i-1} \rangle$   
 $\langle \rangle$  if  $o = 0$  and  $\langle x_{i-1} \rangle$  if  $o = 1$
- ▶  $h_i$  uniquely identifies the  $i$ th shortened random history

How many cpds can we have?



# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\blacktriangleright \langle \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS



# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence}, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Bigram LM (1-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx \langle x_{i-1} \rangle \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence}, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence}, \text{EoS} \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\blacktriangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\blacktriangleright \langle \text{BoS}, \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS



# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, X_5, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence}, X_6 \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Trigram LM (2-order assumption) - illustration

$$N \sim P_N$$

$$\triangleright \langle \text{BoS}, \text{BoS}, X_1, X_2, X_3, X_4, X_5, X_6 = \text{EoS} \rangle$$

$$X_i | x_{<i} \approx x_{i-2}^{i-1} \sim P_{X|H} \text{ for } i = 1, \dots, 6$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, X_2, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, X_3, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, X_4, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, X_5, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence}, X_6 \rangle$$

$$\triangleright \langle \text{BoS}, \text{BoS}, \text{this}, \text{is}, \text{a}, \text{short}, \text{sentence}, \text{EoS} \rangle$$

---

More systematic with BoS and EoS padding:  $o$  leading BoS and 1 trailing EoS

# Factorisation of $n$ -gram LMs

Unigram LM – 0-order Markov model (MM)

the winged irritator challenger<sub>i</sub> pounced upon the hapless bambiraptor

Bigram LM – 1st order MM

1. the | BoS
2. winged | the
3. irritator | winged
4. challenger<sub>i</sub> | irritator
5. pounced | challenger<sub>i</sub>
6. upon | pounced
7. the | upon
8. hapless | the
9. bambiraptor | hapless
10. EoS | bambiraptor

Trigram LM – 2nd order MM

1. the | BoS BoS
2. winged | BoS the
3. irritator | the winged
4. challenger<sub>i</sub> | winged irritator
5. pounced | irritator challenger<sub>i</sub>
6. upon | challenger<sub>i</sub> pounced
7. the | pounced upon
8. hapless | upon the
9. bambiraptor | the hapless
10. EoS | hapless bambiraptor

# MLE for $n$ -gram language models

0th order factorisation and categorical parameterisation

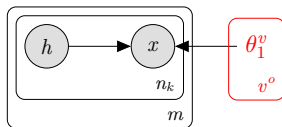
$$\begin{aligned}P_S(\langle x_1, \dots, x_n \rangle) &= P_N(n) \prod_{i=1}^n P_{X|H}(x_i | x_{<i}) \\&\approx c \prod_{i=1}^n P_{X|H}(x_i | x_{i-o}^{i-1}) \\&\propto \prod_{i=1}^n \text{Cat}(X = x_i | \theta_1^{(h_i)}, \dots, \theta_v^{(h_i)})\end{aligned}$$

# MLE for $n$ -gram language models

oth order factorisation and categorical parameterisation

$$\begin{aligned}
 P_S(\langle x_1, \dots, x_n \rangle) &= P_N(n) \prod_{i=1}^n P_{X|H}(x_i | x_{<i}) \\
 &\approx c \prod_{i=1}^n P_{X|H}(x_i | x_{i-o}^{i-1}) \\
 &\propto \prod_{i=1}^n \text{Cat}(X = x_i | \theta_1^{(h_i)}, \dots, \theta_v^{(h_i)})
 \end{aligned}$$

What's the MLE solution for  $\theta_x^{(h)}$ ?

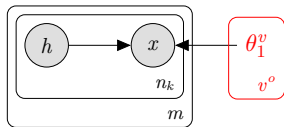


# MLE for $n$ -gram language models

oth order factorisation and categorical parameterisation

$$\begin{aligned}
 P_S(\langle x_1, \dots, x_n \rangle) &= P_N(n) \prod_{i=1}^n P_{X|H}(x_i | x_{<i}) \\
 &\approx c \prod_{i=1}^n P_{X|H}(x_i | x_{i-o}^{i-1}) \\
 &\propto \prod_{i=1}^n \text{Cat}(X = x_i | \theta_1^{(h_i)}, \dots, \theta_v^{(h_i)})
 \end{aligned}$$

What's the MLE solution for  $\theta_x^{(h)}$ ?



$$\begin{aligned}
 \theta_x^{(h)} &= \frac{\sum_{k=1}^m \sum_{i=1}^{n_k} [\textcolor{teal}{h} = h_i^{(k)} \wedge \textcolor{blue}{x} = x_i^{(k)}]}{\sum_{k=1}^m \sum_{i=1}^{n_k} [\textcolor{teal}{h} = h_i^{(k)}]} \\
 &= \frac{\text{count}(\textcolor{teal}{h} \circ \langle \textcolor{blue}{x} \rangle)}{\text{count}(\textcolor{teal}{h})}
 \end{aligned}$$

# NLMI

Probability of a sentence

Language models

Smoothing

Evaluating language models



# Have we really beaten data sparsity?

Dinosaurs have been long extinct

*the winged irritator challengeri pounced upon the hapless bambiraptor*

- ▶ How many of these words do you expect to find in newswire corpora?
- ▶ What about higher-order  $n$ -grams?
- ▶ The probability of the sentence is likely 0
  - ▶ it takes one **unseen**  $n$ -gram  
e.g. *challengeri* or *bambiraptor*

# Have we really beaten data sparsity?

Dinosaurs have been long extinct

*the winged irritator challengeri pounced upon the hapless bambiraptor*

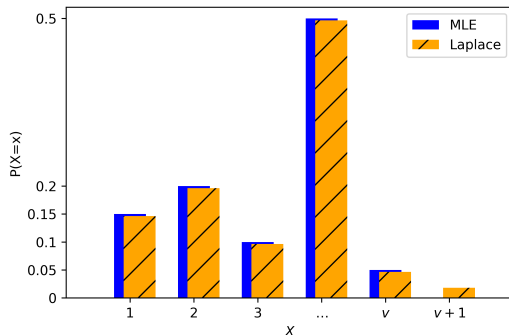
- ▶ How many of these words do you expect to find in newswire corpora?
- ▶ What about higher-order  $n$ -grams?
- ▶ The probability of the sentence is likely 0
  - ▶ it takes one **unseen**  $n$ -gram  
e.g. *challengeri* or *bambiraptor*

MLE assigns probability to **observed**  $n$ -grams only

# Smoothing — Rationale

We can take probability mass away from **seen**  $n$ -grams and reserve such mass to **unseen**  $n$ -grams

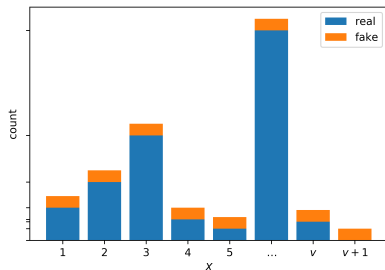
Example unigram distribution



# Laplace smoothing: “add 1 smoothing”

1. Sample space: redefine to include an UNK token  
 $\Sigma := \Sigma \cup \{\text{UNK}\}$       and       $v := v + 1$
2. RV: redefine  $X$  to map unseen symbols to UNK's index
3. MLE: augment counts by 1  
 $\text{count}(\text{UNK}) = 0 \rightarrow 1$

## Example    unigram counts



# Laplace smoothing: “add $\alpha$ smoothing”

Unsmoothed MLE

$$\theta_x^{(h)} = \frac{\text{count}(h, x)}{\text{count}(h)}$$

Laplace-smoothed:  $\alpha > 0$

$$\theta_x^{(h)} = \frac{\text{count}(h, x) + \alpha}{\sum_{x \in \mathcal{X}} \text{count}(h, x) + \alpha} = \frac{\text{count}(h, x) + \alpha}{\text{count}(h) + v\alpha}$$

---

In the lab you will work with another discounting technique:

Good-Turing

# Interpolation

Back-off to further shortened histories

- ▶ if we haven't seen *challengeri pounced upon*  
we might have seen *pounced upon*

Trigram example

$$\begin{aligned}P_{X|H}(x_i|\langle x_{i-2}, x_{i-1} \rangle) &= \lambda_1 P_{X|H}(x_i|\langle x_{i-2}, x_{i-1} \rangle) && \text{trigram} \\&+ \lambda_2 P_{X|H}(x_i|\langle x_{i-1} \rangle) && \text{bigram} \\&+ \lambda_3 P_{X|H}(x_i) && \text{unigram} \\&\text{with } \lambda_1 + \lambda_2 + \lambda_3 = 1\end{aligned}$$

Weights may be a function of history

e.g.  $\lambda_1 := \lambda_1(\langle x_{i-2}, x_{i-1} \rangle)$

# Data pre-processing

Map infrequent types to UNK

- ▶ for example, all types that occur once

This also has the effect of reducing the number of parameters

- ▶ you can always use this to reduce memory requirements  
*including in lab exercises ;)*
- ▶ but use sensible thresholds given the size of the data  
e.g. 1, 2, or 3
- ▶ and always explain your choices

# Bag of tricks

## Smoothing techniques

e.g. discounting, interpolation, data pre-processing

- ▶ are tricks to make MLE more useful
- ▶ some are justified by frequentist statistics
- ▶ some are simply necessary hacks

Manning and Schütze (1999) as well as Jurafsky and Martin (2000) discuss more techniques and in greater detail



# NLMI

Probability of a sentence

Language models

Smoothing

Evaluating language models

# How do we compare language models?

Model  $\mathcal{M}$

- ▶ a set of conditional dependence statements  
graphical structure
- ▶ a parameterisation and a set of parameters  $\theta$

Intrinsic evaluation setup

- ▶ training data: sentences used to estimate parameters
- ▶ test data: a disjoint set of sentences used to assess models

We would like to compare models by comparing

- ▶ the probability they assign to held-out (iid) data  $\mathcal{D}$

$$\prod_{x_1^n \in \mathcal{D}} P_S(x_1^n | \mathcal{M}_1) \stackrel{?}{>} \prod_{x_1^n \in \mathcal{D}} P_S(x_1^n | \mathcal{M}_2)$$

- ▶ but  $P_S$  depends on choice of factorisation

## Perplexity

For dataset  $\mathcal{D}$  and model  $\mathcal{M}$

$$\text{PP}(\mathcal{D}; \mathcal{M}) = \left( \prod_{x_1^n \in \mathcal{D}} P_S(x_1^n; \mathcal{M}) \right)^{-1/t}$$

where  $t$  is the number of tokens in  $\mathcal{D}$

Or in log-domain:  $\log \text{PP}(\mathcal{D}; \mathcal{M}) =$

$$\begin{aligned} &= -\frac{1}{t} \left[ \sum_{k=1}^m \log P_N(n_k) + \log P_{S|N} \left( \langle x_1^{(k)}, \dots, x_{n_k}^{(k)} \rangle | n_k; \mathcal{M} \right) \right] \\ &= -\frac{1}{t} \left[ \log P_{S|N} \left( \langle x_1^{(k)}, \dots, x_{n_k}^{(k)} \rangle | n_k; \mathcal{M} \right) \right] + C \\ &\propto -\frac{1}{t} \left[ \sum_{k=1}^m \sum_{i=1}^{n_k} P_{X|H} \left( x_i^{(k)} | x_{<i}^{(k)}; \mathcal{M} \right) \right] \end{aligned}$$

---

Assuming the length component is the same for every model in the comparison

# Perplexity: interpretation

Perplexity can be seen as

- ▶ *average branching factor* of the language according to the estimated model
- ▶ branching factor: number of words that may follow any word

Comparing models using perplexity require

- ▶ their support must overlap  
i.e. there is a common set of sentences to which both models assign non-zero probability
- ▶ test sentences must be in that common set  
for  $n$ -gram models this typically requires
  - ▶ smoothing
  - ▶ shared vocabulary

# References I

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 1 edition, 2000.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

H. S. Sichel. On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society. Series A (General)*, 137(1):25–34, 1974. ISSN 00359238. URL <http://www.jstor.org/stable/2345142>.