Till Training

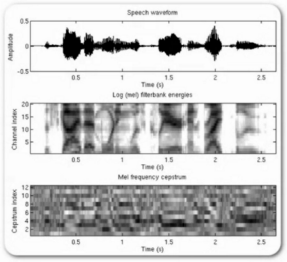
16 April 2021 02:49 PM

Celebrity Sighting

4 / 99

Front End Processing

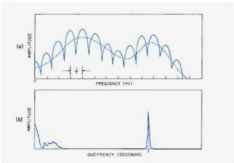
- Convert waveform to *features*.

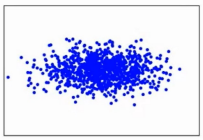


8 / 99

What Have We Gained?

- Time domain ⇒ frequency domain.
- Removed vocal-fold excitation.
- Made features independent.

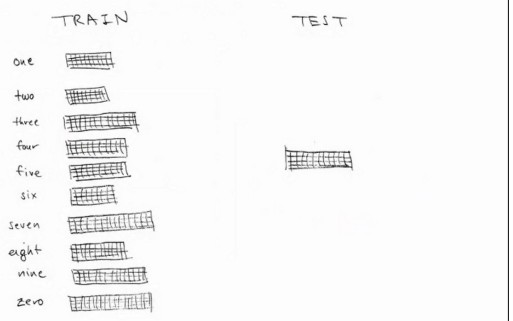




9 / 99

Cepstral Analysis : formant's envelopment after FFT

ASR 1.0: Dynamic Time Warping

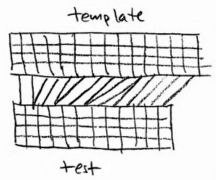


10 / 99

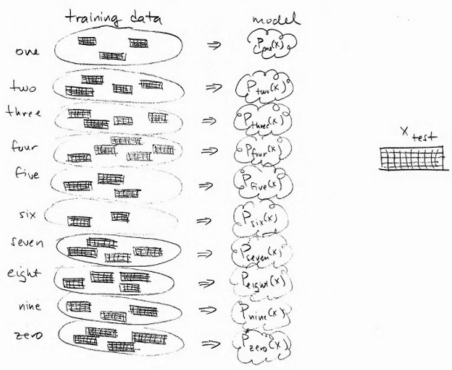
Training Template's
↓
Frequency Domain

Computing the Distance Between Utterances

- Find "best" alignment between frames.
- Sum distances between aligned frames.
- Sum penalties for "weird" alignments.



ASR 2.0: The HMM/GMM Framework



Combination of word → sound's
→ represent it as a Gaussian state
in HMM

Notation

	notation	picture
feature vector	x	
word	ω	seven
model	$P_\omega(x)$	$P_{seven}(x)$

How Do We Do Recognition?

- $\mathbf{x}_{\text{test}}$ = test features; $P_\omega(\mathbf{x})$ = word model.

$$(\text{answer}) = ???$$

$$(\text{answer}) = \arg\max_{\omega \in \text{vocab}} P_\omega(\mathbf{x}_{\text{test}})$$

- Return the word whose model ...
- Assigns the highest prob to the utterance.

Putting it All Together

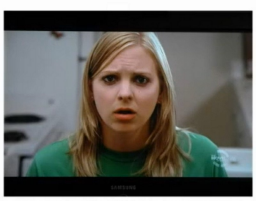
- $P_\omega(\mathbf{x})$ = ???
- How do we actually train?
- How do we actually *decode*?

It's a puzzlement by **jubgo**. Some rights reserved.

So What's the Model?

$$P_\omega(\mathbf{x}) = ???$$

- Frequency that word ω generates features $\mathbf{x}$.
- Has something to do with HMM's and GMM's.



Untitled by **Daniel Olnes**. Some rights reserved.

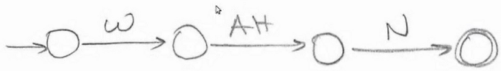
A Word Is A Sequence of Sounds

- *e.g.*, the word ONE: W → AH → N.
- Phoneme inventory.

AA	AE	AH	AO	AW	AX	AXR	AY	B
BD	CH	D	DD	DH	DX	EH	ER	EY
F	G	GD	HH	IH	IX	IY	JH	K
KD	L	M	N	NG	OW	OY	P	PD
R	S	SH	T	TD	TH	TS	UH	UW
V	W	X	Y	Z	ZH			

- What sounds make up TWO?
- What do we use to model sequences?

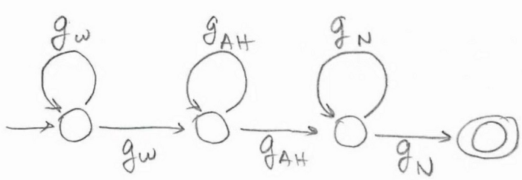
HMM, v1.0



- Outputs on arcs, not states.
- What's the problem? What are the outputs?

What frame corresponds to what frame?

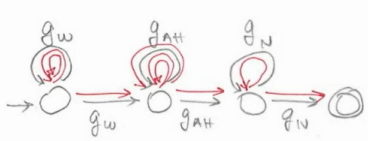
HMM, v3.0



- Are we done?

How long to be in a particular state?
How can I align dragged phones?!

Concept: Alignment ⇔ Path



We keep it in the same state, if it is a dragged phone.
Ideally 1 phone → 1 frame. But not possible due to duration, articulation.
Therefore, we should capture

- Path through HMM ⇒ sequence of arcs, one per frame.
- Notation: $A = a_1 \cdots a_T$.
- a_t = which arc generated frame t.

The Game Plan

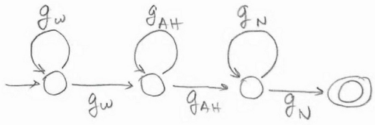
- Express $P_\omega(\mathbf{x})$, the total prob of $\mathbf{x}$ ...
- In terms of $P_\omega(\mathbf{x}, A)$, the prob of a single path.
- How?

$$P(\mathbf{x}) = \sum_{\text{paths } A} (\text{path prob})$$
$$= \sum_{\text{paths } A} P(\mathbf{x}, A)$$

- Sum over all paths.

How To Compute the Likelihood of a Path?

- Path: $A = a_1 \cdots a_T$.



$$P(\mathbf{x}, A) = \prod_{t=1}^{T} (\text{arc prob}) \times (\text{output prob})$$
$$= \prod_{t=1}^{T} p_{a_t} \times P(\vec{x}_t | a_t)$$

- Multiply arc, output probs along path.

What Do Output Probabilities Look Like?

- Mixture of diagonal-covariance Gaussians.

$$P(\vec{x}|a) = \sum_{\text{comp } j} (\text{mixture wgt}) \prod_{\text{dim } d} (\text{Gaussian for dim } d)$$
$$= \sum_{\text{comp } j} p_{a,j} \prod_{\text{dim } d} \mathcal{N}(x_d; \mu_{a,j,d}, \sigma_{a,j,d})$$

articulation.
Therefore, we should capture frame information rather than just utterances.

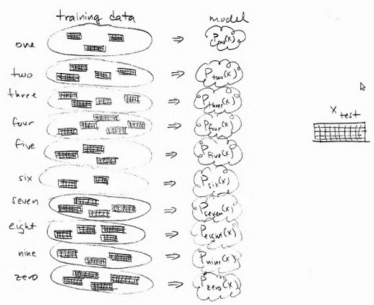
The Full Model

$$
\begin{aligned}
P(\mathbf{x}) &= \sum_{\text{paths } A} P(\mathbf{x}, A) \\
&= \sum_{\text{paths } A} \prod_{t=1}^{T} p_{a_t} \times P(\vec{x}_t | a_t) \\
&= \sum_{\text{paths } A} \prod_{t=1}^{T} p_{a_t} \sum_{\text{comp } j} p_{a_t,j} \prod_{\text{dim } d} \mathcal{N}(x_{t,d}; \mu_{a_t,j,d}, \sigma^2_{a_t,j,d})
\end{aligned}
$$

- p_a — transition probability for arc a.
- $p_{a,j}$ — mixture weight, jth component of GMM on arc a.
- $\mu_{a,j,d}$ — mean, dth dim, jth component, GMM on arc a.
- $\sigma^2_{a,j,d}$ — variance, dth dim, jth component, GMM on arc a.

Training

- How to create model $P_\omega(\mathbf{x})$ from examples $\mathbf{x}_{\omega,1}$, $\mathbf{x}_{\omega,2}$, ...?



What is the Goal of Training?

- To estimate *parameters* ...
- To maximize likelihood of training data.

Crossfit 0303 by Runar Ellertsen. Some rights reserved.

Warm-Up: Non-Hidden ML Estimation

- *e.g.*, Gaussian estimation, non-hidden Markov Models.
- How to do this? (Hint: ??? and ???.)

parameter	description	statistic
p_a	arc prob	# times arc taken
$p_{a,j}$	mixture wgt	# times component used
$\mu_{a,j,d}$	mean	x_d
$\sigma^2_{a,j,d}$	variance	x_d^2

- Count and normalize.
  - *i.e.*, collect a statistic; divide by normalizer count.

How To Estimate Hidden Models?

- The EM algorithm ⇒ FB algorithm for HMM's.
- Hill-climbing maximum-likelihood estimation.

The EM Algorithm

- Expectation step.
  - Using current model, compute posterior counts . . .
  - Prob that thing occurred at time t.
- Maximization step.
  - Like non-hidden MLE, except . . .
  - Use fractional posterior counts instead of whole counts.
- Repeat.

E step: Calculating Posterior Counts

- *e.g.*, posterior count $\gamma(a, t)$ of taking arc a at time t.

$$\begin{aligned}\gamma(a,t) &= \frac{P(\text{paths with arc } a \text{ at time } t)}{P(\text{all paths})} \\ &= \frac{1}{P(\mathbf{x})} \times P(\text{paths from start to src}(a)) \times \\ &\quad P(\text{arc } a \text{ at time } t) \times P(\text{paths from dst}(a) \text{ to end}) \\ &= \frac{1}{P(\mathbf{x})} \times \alpha(\text{src}(a), t-1) \times p_a \times P(\vec{x}_t | a) \times \beta(\text{dst}(a), t)\end{aligned}$$

- Do Forward algorithm: $\alpha(S, t)$, $P(\mathbf{x})$.
- Do Backward algorithm: $\beta(S, t)$.
- Read off posterior counts.

M step: Non-Hidden ML Estimation

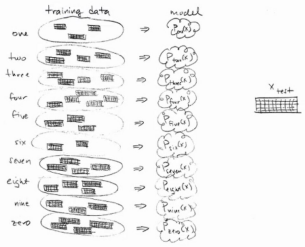
- Count and normalize.
- Same stats as non-hidden, except normalizer is fractional.
- *e.g.*, arc prob p_a

$$p_a = \frac{(\text{count of } a)}{\sum_{\text{src}(a')=\text{src}(a)} (\text{count of } a')} = \frac{\sum_t \gamma(a,t)}{\sum_{\text{src}(a')=\text{src}(a)} \sum_t \gamma(a',t)}$$

- *e.g.*, single Gaussian, mean $\mu_{a,d}$ for dim d.

$$\mu_{a,d} = (\text{mean weighted by } \gamma(a,t)) = \frac{\sum_t \gamma(a,t) x_{t,d}}{\sum_t \gamma(a,t)}$$

What is Decoding?



$$(\text{answer}) = \arg\max_{\omega \in \text{vocab}} P_\omega(\mathbf{x}_{\text{test}})$$

PART II Continuous Word ASR

16 April 2021 03:26 PM

Part II

Continuous Word ASR

Where Are We?

1. Single Word To Continuous Word ASR
2. One More Thing
3. Discussion

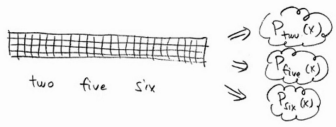
Next Step: Isolated Word Recognition

- It's . . . when . . . you . . . talk . . . like . . . this.

Speech is continuous. Each word flows into another. Not isolated words!

Training Data, Test Data

- What you need: silence-based segmenter.
- Chop up training data.

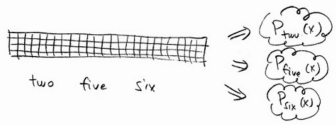


- Chop up test data.
- Reduces to single-word ASR.

No gaps bwn each word.
Cannot segment properly due to continuous nature of phones!

Training Data, Test Data

- What you need: silence-based segmenter.
- Chop up training data.



- Chop up test data.
- Reduces to single-word ASR.

No gaps bwn each word.
Cannot segment properly due to continuous nature of phones!

Continuous Word Recognition

- Single word.
  - Find word model with highest prob:

$$\arg\max_{\omega \in \text{vocab}} P_\omega(\mathbf{x}_{\text{test}})$$

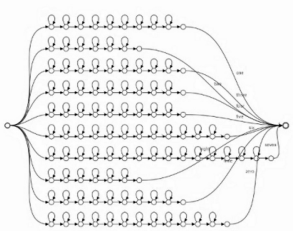
  - V-way classification.
- Continuous word.
  - Find word *sequence* with highest prob:

$$\arg\max_{\omega \in \text{vocab}^*} P(\mathbf{x}_{\text{test}} | \omega)$$

  - ∞-way classification.
- This sounds hard.

One Big HMM Paradigm

- How to modify HMM . . .
  - To accept word sequences instead of single words?

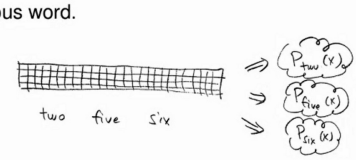


- Loop!

Training on Continuous Word Data

- Isolated word.

- Continuous word.

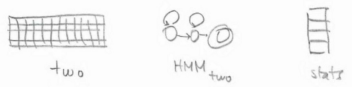


- Don't know where words begin and end!

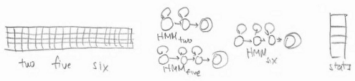
We cannot determine where the phone/word specifically starts or ends.

How Does Training Work Again?

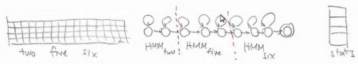
- Isolated word.



- Continuous word: what to do?



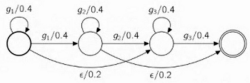
- Idea: concatenate HMM's!



Thus, we need to model silences.

What To Do About Silence?

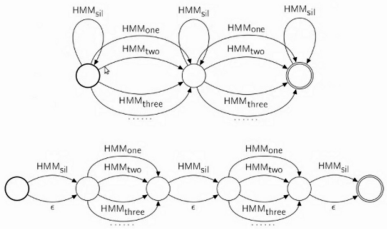
- Treat silence as just another word (~SIL).
- How to design HMM for silence?



Enter WFST

Silence In Decoding

- Where *may* silence occur?
- How many silences can occur in a row?
- Rule of thumb: unnecessary freedom should be avoided.
  - cf. Patriot Act.



$\epsilon \rightarrow$ Unknown state

Word level shown but actually Phone level

Recap: Silence

- Don't forget about silence!
- Silence can be modeled as just another word.
- Generalization: noises, music, filled pauses.

Silence by Alberto Ortiz. Some rights reserved.

HMM/GMM Systems Are Easy To Use

- List of inputs.
- Hyperparameters.
  - HMM topology, # states; # GMM components.
- What else?*
- Utterances with transcripts.
  - Automatically induces word begin/ends, silences.
- Period.

*Small vocabulary only.

Input is GMM alignments.
What frame to what frame a given word is present.

Valid for small vocabularies.

How to model periods.

HMM/GMM Systems Are Flexible

- Same algorithms for:
  - Single word, isolated word, continuous word ASR.
- Just change how HMM is created!

HMM/GMM Systems Are Scalable

- As training data, vocabulary grows.
- In decoding speed.
  - Pruning $\Rightarrow$ time grows slowly.*
- In model size.
  - Number of parameters grow slowly.*

*When using large-vocab methods described in next few lectures.

HMM/GMM's Are The Bomb

- State of art since invented in 1980's.
  - That's 30+ years!
- Until a couple years ago . . .
  - Basically every production system was HMM/GMM.
  - Most probably still are.

BOMB by Aplomb. Some rights reserved.

Obviously these were acoustic models.

Segue: What Have We Learned So Far?

- Small-vocabulary continuous speech recognition.
- What's this good for?
  - Digit strings.
  - Not much else.
- What's next: large-vocabulary CSR.

Part III

Language Modeling

What's the Point?

- ASR works better if you say something "expected".
- Otherwise, it doesn't do that well.

Wreck a Nice Peach?

- Demo.

  THIS IS OUR ROOM FOR A FOUR HOUR PERIOD .
  THIS IS HOUR ROOM FOUR A FOR OUR . PERIOD

  IT IS EASY TO RECOGNIZE SPEECH .
  IT IS EASY TO WRECK A NICE PEACH .
- Homophones; acoustically ambiguous speech.
- How does it get it right . . .
  - Even though acoustics for pair is same?
  - (What if want other member of pair?)
- Need to model "expected" word sequences!

Fails often when encountering OOV utterances :)

Acoustic model captures sound units only.
We also need context of our language!

How Do We Do Recognition?

- $\mathbf{x}_{\text{test}}$ = test features; $P(\mathbf{x}|\omega)$ = HMM/GMM model.

$$(\text{answer}) = ???$$

$$(\text{answer}) = \arg\max_{\omega \in \text{vocab}^*} P(\mathbf{x}_{\text{test}}|\omega)$$

- Return the word sequence that . . .
- Assigns the highest prob to the utterance.

Does This Prefer Likely Word Sequences?

- *e.g.*, $P(\mathbf{x}_{\text{test}}|\text{OUR ROOM})$ *vs.* $P(\mathbf{x}_{\text{test}}|\text{HOUR ROOM})$.
- If I say AA R R UW M, how do these compare?
- They should be about the same.

X_{test} given a bi-gram i.e given context.
This should be learnt by LM

How Do We Fix This?

- Want term $P(\omega)$.
  - Prior over word sequences; prefers likely sequences.
- What HMM/GMM's give us: $P(\mathbf{x}|\omega)$.
- Old: word sequence that maximizes likelihood of feats.

$$(\text{answer}) = \arg\max_{\omega} \; P(\mathbf{x}|\omega)$$

- Idea: most likely word sequence *given* feats!?

$$(\text{answer}) = \arg\max_{\omega} \; P(\omega|\mathbf{x})$$

Bayes' Rule

- The rule:

$$P(\mathbf{x}, \omega) = P(\omega)P(\mathbf{x}|\omega) = P(\mathbf{x})P(\omega|\mathbf{x})$$

$$P(\omega|\mathbf{x}) = \frac{P(\omega)P(\mathbf{x}|\omega)}{P(\mathbf{x})}$$

- Substituting:

$$\begin{aligned}(\text{answer}) &= \arg\max_{\omega}\ P(\omega|\mathbf{x}) \\ &= \arg\max_{\omega}\ \frac{P(\omega)P(\mathbf{x}|\omega)}{P(\mathbf{x})} \\ &= \arg\max_{\omega}\ P(\omega)P(\mathbf{x}|\omega)\end{aligned}$$

P(x|) → Independent

Language Model

Acoustic Model

Priors of word sequences.

The Fundamental Equation of ASR

- Old way:

$$(\text{answer}) = \arg\max_{\omega}\ P(\mathbf{x}|\omega)$$

- New way:

$$(\text{answer}) = \arg\max_{\omega}\ P(\omega|\mathbf{x}) = \arg\max_{\omega}\ P(\omega)P(\mathbf{x}|\omega)$$

- Added term $P(\omega)$, just like we wanted.

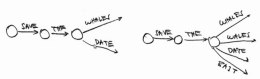
Does This Fix Our Problem?

$$\begin{aligned}(\text{answer}) &= \arg\max_{\omega}\ (\text{language model}) \times (\text{acoustic model}) \\ &= \arg\max_{\omega}\ P(\omega)P(\mathbf{x}|\omega)\end{aligned}$$

- What about homophones?
  THIS IS OUR ROOM FOR A FOUR HOUR PERIOD .
  THIS IS HOUR ROOM FOUR A FOR OUR . PERIOD
- What about confusable sequences in general?
  IT IS EASY TO RECOGNIZE SPEECH .
  IT IS EASY TO WRECK A NICE PEACH .

Language Modeling: Goals

- Describe which word sequences are *likely*.
- Eliminate nonsense; restrict choices given to AM.
  - The fewer choices, the better you do!
- Save acoustic model's ass.



Where Are We?

1. *N*-Gram Models
2. Smoothing
3. How To Smooth
4. Evaluation Metrics
5. Discussion

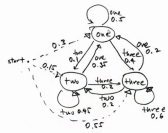
Let's Design a Language Model!

- Goal: probability distribution over word sequences.

$$P(\omega) = P(w_1 w_2 \cdots)$$

- What type of model? What's the simplest we can do?

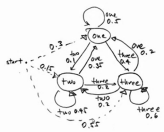
Markov Model, Order 1: Bigram Model



- State ⇔ last word.
- Sum of arc probs leaving state is 1.

0.6, 0.8 self-state's denotes stretching of the words/phones

Bigram Model Example



$P(\text{one three two}) = ???$

$P(\text{one three two}) = 0.3 \times 0.4 \times 0.2 = 0.024$

Bigram Model Equations

$$P(\text{one three two}) = P(\text{one}) \times P(\text{three}|\text{one}) \times P(\text{two}|\text{three})$$
$$= 0.3 \times 0.4 \times 0.2 = 0.024$$

$$P(w_1, \ldots, w_L) = \prod_{i=1}^{L} P(\text{cur word} \mid \text{last word})$$
$$= \prod_{i=1}^{L} P(w_i | w_{i-1})$$

What Training Data?

- Text! As a list of *utterances*.

  I WANT TO FLY FROM AUSTIN TO BOSTON
  CAN I GET A VEGETARIAN MEAL
  DO YOU HAVE ANYTHING THAT IS NONSTOP
  I WANT TO LEAVE ON FEBRUARY TWENTY SEVEN
  WHO LET THE DOGS OUT
  GIVE ME A ONE WAY TICKET PAUSE TO HELL

- Are AM's or LM's usually trained with more data?

Language Models are mainly responsible for addressing homophone issues & incomplete utterances.

Acoustic Model responsible mainly for sound's (featu

LM obviously is responsible for words. pauses, elongate

Incomplete Utterances

- Example: I'M GOING TO

  $P(\text{I'M}) \times P(\text{GOING}|\text{I'M}) \times P(\text{TO}|\text{GOING})$

- Is this a good utterance?
- Does this get a good score?
- How to fix this?

Incomplete image by SamSach. Some rights reserved.

Like CLS & SEP in BERT Lol.

Utterance Begins and Ends

- Add beginning-of-sentence token; *i.e.*, $w_0 = \triangleright$.
- Predict end-of-sentence token at end; *i.e.*, $w_{L+1} = \triangleleft$.

$$P(w_1 \cdots w_L) = \prod_{i=1}^{L+1} P(w_i | w_{i-1})$$

- Does this fix problem?

$$P(\text{I'M GOING TO}) = P(\text{I'M}|\triangleright) \times P(\text{GOING}|\text{I'M}) \times P(\text{TO}|\text{GOING}) \times P(\triangleleft|\text{TO})$$

*Side effect: $\sum_\omega P(\omega) = 1$. (Can you prove this?)

How to Set Probabilities?

- How to estimate $P(\text{FLY}|\text{TO})$?

$$P(\text{FLY}|\text{TO}) = \frac{\text{count}(\text{TO FLY})}{\text{count}(\text{TO})}$$

- MLE: count and normalize!

$$P_{\text{MLE}}(w_i | w_{i-1}) = \frac{c(w_{i-1} w_i)}{\sum_w c(w_{i-1} w)}$$
$$= \frac{c(w_{i-1} w_i)}{c(w_{i-1})}$$

P(X) scores of words in entire Vocabulary.

wrong)

ed words

Example: Maximum Likelihood Estimation

- 23M words of Wall Street Journal text.

  FEDERAL HOME LOAN MORTGAGE CORPORATION –DASH ONE .POINT FIVE BILLION DOLLARS OF REALESTATE MORTGAGE -HYPHEN INVESTMENT CONDUIT SECURITIES OFFERED BY MERRILL LYNCH &AMPERSAND COMPANY .PERIOD

  NONCOMPETITIVE TENDERS MUST BE RECEIVED BY NOON EASTERN TIME THURSDAY AT THE TREASURY OR AT FEDERAL RESERVE BANKS OR BRANCHES .PERIOD

  THE PROGRAM ,COMMA USING SONG ,COMMA PUPPETS AND VIDEO ,COMMA WAS CREATED AT LAWRENCE LIVERMORE NATIONAL LABORATORY ,COMMA LIVERMORE ,COMMA CALIFORNIA ,COMMA AFTER A PARENT AT A BERKELEY ELEMENTARY SCHOOL EXPRESSED INTEREST .PERIOD

Example: Bigram Model

- P(I HATE TO WAIT) $=$???
- P(EYE HATE TWO WEIGHT) $=$???
- Step 1: Collect all bigram counts, unigram history counts.

	EYE	I	HATE	TO	TWO	WAIT	WEIGHT	$\triangleleft$	$*$
$\triangleright$	3	3234	5	4064	1339	8	22	0	892669
EYE	0	0	0	26	1	0	0	52	735
I	0	0	45	2	1	1	0	8	21891
HATE	0	0	0	40	0	0	0	9	246
TO	8	6	19	21	5341	324	4	221	510508
TWO	0	5	0	1617	652	0	0	4213	132914
WAIT	0	0	0	71	2	0	0	35	882
WEIGHT	0	0	0	38	0	0	0	45	643