A famous quote

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

Intuitive interpretation

- “Probability of a sentence” = how likely is it to occur in natural language
  - Consider only a specific language (English)
  - Not including meta-language (e.g. linguistic discussion)

\[ P(\text{She studies morphosyntax}) > P(\text{She studies more faux syntax}) \]
**Automatic speech recognition**

Sentence probabilities (language model) help decide between similar-sounding options.

speech input

↓ (Acoustic model)

possible outputs

↓ (Language model)

best-guess output

She studies morphosyntax
She studies more faux syntax
She’s studies morph or syntax
...

↓ (Language model)

best-guess output

She studies morphosyntax

---

**So, not “entirely useless”...**

- Sentence probabilities are clearly useful for language engineering [this course].
- Given time, I could argue why they’re also useful in linguistic science (e.g., psycholinguistics). But that’s another course...

---

**Machine translation**

Sentence probabilities help decide word choice and word order.

non-English input

↓ (Translation model)

possible outputs

↓ (Language model)

best-guess output

She is going home
She is going house
She is traveling to home
To home she is going
...

↓ (Language model)

best-guess output

She is going home

---

**But, what about zero probability sentences?**

the Archaeopteryx winged jaggedly amidst foliage
vs
jaggedly trees the on flew

- Neither has ever occurred before.
  ⇒ both have zero probability.
- But one is grammatical (and meaningful), the other not.
  ⇒ “Sentence probability” is useless as a measure of grammaticality.
The logical flaw

• “Probability of a sentence” = how likely is it to occur in natural language.

• Is the following statement true?
  
  Sentence has never occurred ⇒ sentence has zero probability

• More generally, is this one?
  
  Event has never occurred ⇒ event has zero probability

Events that have never occurred

• Each of these events has never occurred:
  
  My hair turns blue
  I injure myself in a skiing accident
  I travel to Finland

• Yet, they clearly have different (and non-zero!) probabilities.

Events that have never occurred

• Each of these events has never occurred:
  
  My hair turns blue
  I injure myself in a skiing accident
  I travel to Finland

• Yet, they clearly have differing (and non-zero!) probabilities.

• Most sentences (and events) have never occurred.
  
  – This doesn’t make their probabilities zero (or meaningless), but
  – it does make estimating their probabilities trickier.

Probability theory vs estimation

• Probability theory can solve problems like:
  
  – I have a jar with 6 blue marbles and 4 red ones.
  – If I choose a marble uniformly at random, what’s the probability it’s red?

• But what about:
  
  – I have a jar of marbles.
  – I repeatedly choose a marble uniformly at random and then replace it before choosing again.
  – In ten draws, I get 6 blue marbles and 4 red ones.
  – On the next draw, what’s the probability I get a red marble?

• The latter also requires estimation theory.
Example: weather forecasting

What is the probability that it will rain tomorrow?

- To answer this question, we need
  - data: measurements of relevant info (e.g., humidity, wind speed/direction, temperature).
  - model: equations/procedures to estimate the probability using the data.

- In fact, to build the model, we will need data (including outcomes) from previous situations as well.

Example: language model

What is the probability of sentence $\overrightarrow{w} = w_1 \ldots w_n$?

- To answer this question, we need
  - data: words $w_1 \ldots w_n$, plus a large corpus of sentences ("previous situations", or training data).
  - model: equations to estimate the probability using the data.

- Different models will yield different estimates, even with the same data.

- Deep question: what model/estimation method do humans use?

How to get better probability estimates

Better estimates definitely help in language technology. How to improve them?

- More training data. Limited by time, money. (Varies a lot!)

- Better model. Limited by scientific and mathematical knowledge, computational resources

- Better estimation method. Limited by mathematical knowledge, computational resources

We will return to the question of how to know if estimates are “better”.

Sharon Goldwater ANLP Lecture 4 12

Sharon Goldwater ANLP Lecture 4 13

Sharon Goldwater ANLP Lecture 4 14

Sharon Goldwater ANLP Lecture 4 15
**Notation**

• When the distinction is important, will use
  – \( P(\vec{w}) \) for *true* probabilities
  – \( \hat{P}(\vec{w}) \) for *estimated* probabilities
  – \( P_E(\vec{w}) \) for estimated probabilities using a particular estimation method \( E \).

• But since we almost always mean estimated probabilities, may get lazy later and use \( P(\vec{w}) \) for those too.

---

**Example: estimation for coins**

I flip a coin 10 times, getting 7T, 3H. What is \( \hat{P}(T) \)?

• **Model 1:** Coin is fair. Then, \( \hat{P}(T) = 0.5 \)

• **Model 2:** Coin is not fair. Then, \( \hat{P}(T) = 0.7 \) (why?)
Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is \( \hat{P}(T) \)?

- **Model 1**: Coin is fair. Then, \( \hat{P}(T) = 0.5 \)
- **Model 2**: Coin is not fair. Then, \( \hat{P}(T) = 0.7 \) (why?)
- **Model 3**: Two coins, one fair and one not; choose one at random to flip 10 times. Then, \( 0.5 < \hat{P}(T) < 0.7 \).

Defining a model

Usually, two choices in defining a model:

- **Structure** (or form) of the model: the form of the equations, usually determined by knowledge about the problem.
- **Parameters** of the model: specific values in the equations that are usually determined using the training data.

Example: height of 30-yr-old females

Assume the form of a normal distribution, with parameters \((\mu, \sigma)\):

\[
p(x|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)
\]
Example: height of 30-yr-old females

Collect data to determine values of $\mu, \sigma$ that fit this particular dataset.

![Height frequency distribution graph](image)

Example: M&M colors

What is the proportion of each color of M&M?

- Assume a **discrete distribution** with parameters $\theta$.
  - $\theta$ is a vector! That is, $\theta = (\theta_R, \theta_O, \theta_Y, \theta_G, \theta_B, \theta_Br)$.
  - For discrete distribution, the parameters ARE the probabilities, e.g., $P(\text{red}) = \theta_R$.

- In 48 packages, I find\(^1\) 2620 M&Ms, as follows:
  
<table>
<thead>
<tr>
<th>Color</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>372</td>
</tr>
<tr>
<td>Orange</td>
<td>544</td>
</tr>
<tr>
<td>Yellow</td>
<td>369</td>
</tr>
<tr>
<td>Green</td>
<td>483</td>
</tr>
<tr>
<td>Blue</td>
<td>481</td>
</tr>
<tr>
<td>Brown</td>
<td>371</td>
</tr>
</tbody>
</table>

- How to estimate $\theta$ from this data?

\(^1\)Actually, data from: https://joshmadison.com/2007/12/02/mms-color-distribution-analysis/

Relative frequency estimation

- Intuitive way to estimate discrete probabilities:

  $$P_{RF}(x) = \frac{C(x)}{N}$$

  where $C(x)$ is the count of $x$ in a large dataset, and $N = \sum_{x'} C(x')$ is the total number of items in the dataset.

- M&M example: $P_{RF}(\text{red}) = \hat{\theta}_{R} = \frac{372}{2620} = .142$

- Or, could estimate probability of word $w$ from a large corpus.

- Can we justify this mathematically?
Formalizing the estimation problem

- What is the best choice of $\theta$ given the data $d$ that we saw?
- Formalize using Bayes’ Rule, try to maximize $P(\theta|d)$.

\[
P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}
\]

- $P(\theta)$: prior probability of $\theta$
- $P(d|\theta)$: likelihood
- $P(\theta|d)$: posterior probability of $\theta$ given $d$

Maximum-likelihood estimation

- Not obvious how to choose prior... so maybe don’t use one?

\[
\hat{\theta} = \arg\max_\theta P(\theta|d)
\]

- Just choose $\theta$ to maximize the likelihood.
  – the parameters that make the observed data most probable
- This turns out to be just the relative frequency estimator, i.e.,

\[
P_{\text{ML}}(x) = P_{\text{RF}}(x) = \frac{C(x)}{N}
\]
**Likelihood example**

- For a fixed dataset, the likelihood depends on the model we use.
- Our coin example: \( \theta = (\theta_H, \theta_T) \). Suppose \( d = \text{HTTTHTHTTT} \).
- **Model 1**: Assume coin is fair, so \( \hat{\theta} = (0.5, 0.5) \).
  - Likelihood of this model:
    \[
    P(\text{HTTTHTHTTT}|\hat{\theta}) = (0.5)^3 \cdot (0.5)^7 = 0.00097
    \]

**Sharon Goldwater ANLP Lecture 4 32**

**Summary**

- “Probability of a sentence”: how likely is it to occur in natural language?
- Useful in many natural language applications (and linguistics)
- Can never know the true probability, but we may be able to estimate it.
- Probability estimates depend on
  - The data we have observed
  - The model (structure and parameters) we choose
- One way to estimate probabilities: maximum-likelihood estimation

**Likelihood example**

- For a fixed dataset, the likelihood depends on the model we use.
- Our coin example: \( \theta = (\theta_H, \theta_T) \). Suppose \( d = \text{HTTTHTHTTT} \).
- **Model 1**: Assume coin is fair, so \( \hat{\theta} = (0.5, 0.5) \).
  - Likelihood of this model:
    \[
    P(\text{HTTTHTHTTT}|\hat{\theta}) = (0.5)^3 \cdot (0.5)^7 = 0.00097
    \]
- **Model 2**: Use ML estimation, so \( \hat{\theta} = (0.3, 0.7) \).
  - Likelihood of this model: \( (0.3)^3 \cdot (0.7)^7 = 0.00222 \)
- Maximum-likelihood estimate does have higher likelihood!

**Where to go from here?**

Next time, we’ll start to discuss

- Different generative models for sentences (model structure), and the questions they can address
- Weaknesses of MLE and ways to address them (parameter estimation methods)
Reminders/Announcements

Tutorial groups tomorrow/Wed:

- Check the web page to see which group to attend this week.