This lecture

- An extractive summary of the course.
Exam

• 15 April from 19h00—22h00.

<table>
<thead>
<tr>
<th>A-Kar</th>
<th>Kas-Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB404</td>
<td>GB405</td>
</tr>
</tbody>
</table>

Galbraith building, 35 St. George St.

• No aids allowed – your desk should have nothing but:
  • Your UofT ID,
  • The exam, and
  • A writing implement.
Structure

• Following the format of previous years:
  • 20 multiple-choice questions [40 marks]
    • 4 options each.
  • 10 short-answer questions [30 marks]
    • Some of these involve simply giving a definition. Others involve some calculation.
  • 3 subject-specific questions [30 marks]
    • These questions involve a small component of original thinking.
Hints for studying

• **Definitions:** *n.pl.* Terms that are useful to know.

  • Highlights are also useful to know.

• Not all definitions/highlights are in the exam.
• Not all things on the exam have been highlighted.
  • This review lecture is not a substitute for the rest of the material in this course.
Hints for studying

• Go through the exams on the website from previous years (and focus on those items related to topics we covered this year).

• Go through the quiz from this year.

• Work out worked-out examples for yourself, ideally more than once.

• I find it helpful to just relax before an exam.
Exam material

• The exam covers all material in the lectures and assignments except:
  • Material in the bonuses of assignments, and  
  • Slides with ‘Aside’ in the title.

• The reading material (e.g., Manning & Schütze) provides background to concepts discussed in class.
  • If a concept appears in the textbook but not in the lectures/assignments, you don’t need to know it, even if it’s very interesting.
Categories of linguistic knowledge

• **Phonology**: the study of patterns of speech sounds.
  e.g., “read” → /r iy d/

• **Morphology**: how words can be changed by inflection or derivation.
  e.g., “read”, “reads”, “reader”, “reading”, ...

• **Syntax**: the ordering and structure between words and phrases.
  e.g., NounPhrase → det. adj. n.

• **Semantics**: the study of how meaning is created by words and phrases.
  e.g., “book” →

• **Pragmatics**: the study of meaning in broad contexts.
Ambiguity – Phonological

- **Phonology**: the study of patterns of speech sounds.

  - “read” → /r iy d/ as in ‘I like to read’
  - “read” → /r eh d/ as in ‘She read a book’
  - “object” → /aa₁ b jh eh⁰ k t / as in ‘That is an object’
  - “object” → /ah⁰ b jh eh¹ k t / as in ‘I object!’
  - “too” ← /t uw/ as in ‘too salty’
  - “two” ← /t uw/ as in ‘two beers’

- Ambiguities can often be resolved in context, but not always.
  - e.g., /h aw t uw r eh¹ k ah ?? n ay² z s (b|p) iy ch/ → ‘how to recognize speech’
  - → ‘how to wreck a nice beach’
Ambiguity – Syntactic

• **Syntax**: the ordering and structure between words. Words can be grouped into ‘parse tree’ structures given grammatical ‘rules’.

  e.g., “I shot an elephant in my pyjamas”
Statistics: what are we counting?

• Almost all statistics are based on simple counting.
• What are we counting?
  
  First, we shape our tools and thereafter our tools shape us.

• Tokens: *n.pl. instances* of words or punctuation (13).

• Types: *n.pl. ‘kinds’* of words or punctuation (10).
Corpora

- **Corpus**: *n.* A body of language data of a particular sort (*pl.* corpora).

- Most **valuable** corpora occur **naturally**
  - e.g., newspaper articles, telephone conversations, multilingual transcripts of the United Nations

- We use corpora to gather statistics; more is better (typically between $10^7$ and $10^{12}$ tokens).
Notable corpora

- **Penn treebank**: Syntactically annotated Brown, plus others incl. 1989 *Wall Street Journal*.
- **Switchboard corpus**: 120 hours \(\approx 2.4\text{M tokens}\). 2.4K telephone conversations between US English speakers.
Additional notable corpora

- **Hansard corpus**: Canadian parliamentary proceedings, French/English bilingual.
- **Gutenberg project**: 33K free eBooks, several languages. [http://www.gutenberg.org](http://www.gutenberg.org)
- **Google corpus**: Index of between $10^{11}$ and $10^{12}$ 5-word sequences (13,588,391 word types (incl. numbers, names, misspellings, etc.)) [http://ngrams.googlelabs.com/](http://ngrams.googlelabs.com/)
Frequency statistics

• **Term frequency** of token $w$ in corpus $C$ is the number of times token $w$ occurs in $C$.

  $Count(w, C)$

• **Relative frequency** is defined relative to the size of the corpus, $\|C\|$ (the number of words in the corpus).

  $$F_C(w) = \frac{Count(w, C)}{\|C\|}$$

• In theory, as $\|C\|$ increases, $F_C(w)$ approaches $P(w)$. This is the frequentist view.
Very simple predictions

• Let’s return to word prediction.
• We want to know the probability of the next word given the previous words in a sequence.

• We can approximate conditional probabilities by counting occurrences in large corpora of data.
  • E.g., \( P(\text{food} \mid I \text{ want Chinese}) = \frac{P(I \text{ want Chinese food})}{P(I \text{ want Chinese})} \approx \frac{\text{Count}(I \text{ want Chinese food})}{\text{Count}(I \text{ want Chinese})} \)

Why? Hint: Corpus size
Bayes’ theorem

\[ P(A, B) = P(A)P(B | A) \]
\[ P(A, B) = P(B)P(A | B) \]

Bayes theorem: \( P(A | B) = \frac{P(B | A)P(A)}{P(B)} \)
Markov assumption

• Assume each observation only depends on a short linear history of length $L$.

\[ P(w_n | w_1:(n-1)) \approx P(w_n | w_{(n-L):(n-1)}) \]

• Bigram version:

\[ P(w_n | w_1:(n-1)) \approx P(w_n | w_{n-1}) \]

• Why do we need to make this assumption?
Maximum likelihood estimate

• Maximum likelihood estimate (MLE) of parameters $\theta$ in a model $M$, given training data $T$ is the estimate that maximizes the likelihood of the training data using the model.

• e.g., $T$ is the Brown corpus, $M$ is the bigram and unigram tables $\theta_{(to|\text{want})}$ is $P(to|\text{want})$.

• In fact, we have been doing MLE all along with our simple counting.
Sparsity

• Problem with $N$-gram models:
  • New words appear often.
    • e.g., *interfrastic*, *espepsia*, 182 321.09
  • New bigrams occur *even more* often.
  • New trigrams occur *even more* even-more-often.
Sparsity of unigrams vs. bigrams

- E.g., we’ve seen lots of every unigram, but are missing many bigrams:

```
<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>want</th>
<th>to</th>
<th>eat</th>
<th>Chinese</th>
<th>food</th>
<th>lunch</th>
<th>spend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>2533</td>
<td>927</td>
<td>2417</td>
<td>746</td>
<td>158</td>
<td>1093</td>
<td>341</td>
<td>278</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>want</th>
<th>to</th>
<th>eat</th>
<th>Chinese</th>
<th>food</th>
<th>lunch</th>
<th>spend</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>5</td>
<td>827</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>want</td>
<td>2</td>
<td>0</td>
<td>608</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>to</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>686</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>211</td>
</tr>
<tr>
<td>eat</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>16</td>
<td>2</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>Chinese</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>82</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>food</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lunch</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>spend</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Zipf’s law

• In *Human Behavior [sic] and the Principle of Least Effort*, Zipf argues that all human endeavour depends on laziness.
  • Speaker minimizes effort by having a **small** vocabulary of **common** words.
  • Hearer minimizes effort by having a **large** vocabulary of **less ambiguous** words.

• Compromise: frequency and rank are inversely proportional.

\[
f \propto \frac{1}{r} \quad \text{i.e., for some } k \quad f \cdot r = k
\]
Zipf’s law on the Brown corpus

From Manning & Schütze
Add-1 smoothing (Laplace)

• Given a vocab size $|\mathcal{V}|$ and corpus size $N$, just add 1 to all the counts! No more zeros!

• MLE
  
  $P(w_i) = \frac{C(w_i)}{N}$

• Laplace estimate
  
  $P_{Lap}(w_i) = \frac{C(w_i) + 1}{N + |\mathcal{V}|}$

• Does this give a proper probability distribution? Yes:
  
  $$\sum_w P_{Lap}(w) = \sum_w \frac{C(w) + 1}{N + |\mathcal{V}|} = \frac{\sum_w C(w) + \sum_w 1}{N + |\mathcal{V}|} = \frac{N + |\mathcal{V}|}{N + |\mathcal{V}|} = 1$$
Add-$\delta$ smoothing

- Laplace’s method generalizes a bit:

- Add-$\delta$ estimate:

$$P_\delta(w_i) = \frac{C(w_i) + \delta}{N + \delta \|V\|}$$
Discounting vs. backoff/interpolation

• **Discounting** (e.g., add-1 smoothing) pretends to see an event fewer times (proportionally!) than it really did.
  • (and pretends to have actually seen unseen events)

• **Backoff/interpolation** essentially decides to use more or less $n$-gram context, depending on the situation.
  • If we have 0 probability for an $n$-gram, we can still use the probability of an $(n-1)$-gram.
  • Remember **Katz** and **Linear interpolation**
Feature vectors

- Values for several features of an observation can be put into a single vector.

<table>
<thead>
<tr>
<th>Tweet 1</th>
<th>Proper Nouns</th>
<th>1st Person Pronouns</th>
<th>Commas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Damien Fahey</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>“Rush Limbaugh looks like if someone put a normal human being in landscape mode.”</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tweet 2</th>
<th>Proper Nouns</th>
<th>1st Person Pronouns</th>
<th>Commas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faux John Madden</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>“BREAKING: Apple Maps projecting Barack Obama to win Brazil.”</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tweet 3</th>
<th>Proper Nouns</th>
<th>1st Person Pronouns</th>
<th>Commas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jim Gaffigan</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>“If there was an award for most pessimistic, I probably wouldn't even be nominated.”</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
mRMR feature selection

• Minimum-redundancy-maximum-relevance (mRMR) can use correlation, distance scores (e.g., $D_{KL}$) or mutual information to select features as in

• For feature set $S$ of features $f_i$, class $c$, $D(S, c)$: a measure of relevance of $S$ for $c$, and

$R(S)$: a measure of the redundancy of $S$,

$$S_{mRMR} = \arg\max_s [D(S, c) - R(S)]$$
Parts of speech (PoS)

• Linguists like to group words according to their **structural function** in building sentences.
  • This is similar to grouping Lego by their shapes.

• **Part-of-speech:** *n.* lexical category or morphological class.

Nouns collectively constitute a part of speech (called *Noun*)
Parts of speech (PoS)

• Things that are useful to know about PoS:
  • **Content words** vs. **function words**
  • **Properties** of content words (e.g., number).
  • **Agreement**. Verbs and nouns have to match in number in English (e.g., “the dogs runs” is wrong.)
  • What **PoS Tagging** is, and perhaps some vague idea of how to do it.
Information and entropy
**Entropy**

- **Entropy**: *n.* the **average** amount of information we get in observing the output of source $S$.

$$H(S) = \sum_i p_i I(w_i) = \sum_i p_i \log_2 \frac{1}{p_i}$$
Joint entropy

- **Joint Entropy**: *n.* the **average** amount of information needed to specify multiple variables simultaneously.

$$H(X, Y) = \sum_x \sum_y p(x, y) \log_2 \frac{1}{p(x, y)}$$

Same general form as entropy, except you sum over each variable, and probabilities are joint.
Conditional entropy

- **Conditional entropy**: \( n \). the *average* amount of information needed to specify one variable *given* that you know another.

\[
H(Y|X) = \sum_{x \in X} p(x)H(Y|X = x)
\]

It’s *an average of entropies* over all possible conditioning values.
Mutual information

- Mutual information: *n.* the average amount of information shared between variables.

\[
I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \\
= \sum_{x,y} p(x, y) \log_2 \frac{p(x,y)}{p(x)p(y)}
\]

Again, a sum over each variable, but the log fraction is normalized by an assumption that they’re independent (\(p(x)p(y)\)).
Relations between entropies

\[ H(X, Y) = H(X) + H(Y) - I(X; Y) \]
Information theory

• In general, lecture 2-2 includes some walked-through examples of applying the preceding formula.
• It’s probably a good idea to walk through these examples yourself on paper.
Collocations

• **Collocation**: *n.* a ‘turn-of-phrase’ or usage where a sequence of words is **perceived** to have a meaning ‘**beyond**’ the sum of its parts.

• E.g., ‘*disk drive*’, ‘*video recorder*’, and ‘*soft drink*’ are collocations. ‘*cylinder drive*’, ‘*video measurer*’, ‘*weak drink*’ are **not** despite some near-synonymy between alternatives.

• Collocations are **not** just highly frequent bigrams, otherwise ‘**of the**’, and ‘*and the*’ would be collocations.

• How can we test if a bigram is a collocation or not?
Decision trees

• Consists of *rules* for classifying data that consists of many *attributes*.

• Walk through the Simpsons example from 3-2.
Information gain

• The information gain is based on the expected decrease in entropy after a set of training data is split on an attribute.
• We prefer to choose the attribute that results in more entropy reduction.

\[
Gain(Q) = H(S) - \sum_{\text{child set}} p(\text{child set})H(\text{child set})
\]

\[S = A \cup B\]
\[\emptyset = A \cap B\]
Aspects of ID3

• ID3 tends to build a **short tree** since at each step we are removing the maximum amount of entropy as possible.

• ID3 trains on the **whole training set** and does not succumb to issues related to random initialization.

• ID3 can **over-fit** to training data.

• Only one attribute is used at a time to make decisions

• It can be difficult to use continuous data, since many trees need to be generated to see where to break the continuum.
Markov models
Observable Markov model

- Probabilities on all outgoing arcs must sum to 1.

- \( P(\text{ship}|\text{ship}) + P(\text{tops}|\text{ship}) + P(\text{pass}|\text{ship}) = 1 \)

- \( P(\text{ship}|\text{tops}) + P(\text{tops}|\text{tops}) + P(\text{mother}|\text{tops}) = 1 \)

- ...
Multivariate systems

- What if a conditioning variable changes over time?
  - e.g., I’m happy one second and disgusted the next.
- Here, the **state** is the mood and the **observation** is the word.
Observable multivariate systems

Q: How do you **learn** these probabilities?

- \( P(w_{0:t}, q_{0:t}) \approx \prod_{i=0}^{t} P(q_i | q_{i-1})P(w_i | q_i) \)

A: Basically, the same as before.

- \( P(q_i | q_{i-1}) = \frac{P(q_{i-1} q_i)}{P(q_{i-1})} \) is learned with MLE from training data.
- \( P(w_i | q_i) = \frac{P(w_i q_i)}{P(q_i)} \) is also learned with MLE from training data.
Hidden variables

• Q: What if you don’t have access to the state during testing?
  • e.g., you’re asked to compute $P(\langle ship, ship \rangle)$

• Q: What if you don’t have access to the state during training?
Questions for HMMs

1. Given a model with particular parameters $\theta = \langle \Pi, A, B \rangle$, how do we efficiently compute the likelihood of a particular observation sequence, $P(\mathcal{O}; \theta)$?

2. Given an observation sequence $\mathcal{O}$ and a model $\theta$, how do we choose a state sequence $Q = \{q_0, \ldots, q_T\}$ that best explains the observations?

3. Given a large observation sequence $\mathcal{O}$, how do we choose the best parameters $\theta = \langle \Pi, A, B \rangle$ that explain the data $\mathcal{O}$?
Trellis

State

\[ S_1 \]
\[ S_2 \]
\[ S_3 \]
\[ S_N \]

Time, \( t \)

0
1
2
\( T - 1 \)

Probability of being in state \( s_3 \) at time \( t = 2 \)
0. Initializing the Forward procedure

\[ \alpha_i(0) := \pi_i b_i(\sigma_0), \quad i := 1..N \]

(Probability of starting in state \(i\) and reading the first symbol there)
1. Forward induction

\[ \alpha_j(t + 1) := \sum_{i=1}^{N} \alpha_i(t) a_{ij} b_j(\sigma_{t+1}), \]

\[ j := 1..N, t := 1..(T - 1) \]

(Probability of getting to state \( j \) at time \( t + 1 \))
2. Forward conclusion

\[
P(\mathcal{O}; \theta) = \sum_{i=1}^{N} \alpha_i (T - 1)
\]

Sum over all possible final states.

\[
\text{State: } S_1, S_2, S_3, \ldots, S_N
\]

\[
\text{Time, } t: 0, 1, 2, \ldots, T - 1
\]
Choosing the best state sequence

I want to guess which sequence of states generated an observation.

E.g., if states are PoS and observations are words
The Viterbi algorithm

• Also an inductive dynamic-programming algorithm that uses the trellis.

• Define the probability of the most probable path leading to the trellis node at (state $i$, time $t$) as

$$\delta_i(t) = \max_{q_0 \ldots q_{t-1}} P(q_0 \ldots q_{t-1}, \sigma_0 \ldots \sigma_{t-1}, q_t = s_i; \theta)$$

• And the incoming arc that led to this most probable path is defined as $\psi_i(t)$
Training HMMs

• We want to modify the parameters of our model $\theta = \langle \Pi, A, B \rangle$ so that $P(\mathcal{O}; \theta)$ is maximized for some training data $\mathcal{O}$:

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{O}; \theta)$$

• If we want to choose a best state sequence $Q^*$ on previously unseen test data, the parameters of the HMM should first be tuned to similar training data.
Expectation-maximization

• If we knew $\theta$, we could estimate expectations such as
  • Expected number of times in state $s_i$,  
  • Expected number of transitions $s_i \rightarrow s_j$

• If we knew:
  • Expected number of times in state $s_i$,  
  • Expected number of transitions $s_i \rightarrow s_j$

then we could compute the maximum likelihood estimate of

$$\theta = \langle \pi_i, \{a_{ij}\}, \{b_i(w)\} \rangle$$
Statistical machine translation

STICK ONE IN YOUR EAR, YOU CAN INSTANTLY UNDERSTAND ANYTHING SAID TO YOU IN ANY FORM OF LANGUAGE: THE SPEECH YOU HEAR DECODES THE BRAIN WAVE MATRIX.
Challenges of SMT

- Lexical ambiguity (e.g., words are polysemous).
- Differing word orders.
- Syntactic ambiguity.
- Miscellaneous idiosyncracies.

- Sentence alignment.
  - Gale & Church: alignment by length (minimize costs).
  - Church: cognates approximated by 4-graphs.
  - Melamed: cognates approximated by longest common subsequences.
The noisy channel

Language model

Source

\[ P(E) \]

Decoder

\[ E^* = \arg \max_E P(F|E)P(E) \]

Translation model

Channel

\[ P(F|E) \]

Observed

\[ F' \]
Word alignment

- **Word alignments** can be 1:1, N:1, 1:N, 0:1, 1:0, ... E.g.,

  - "zero fertility" word: not translated (1:0)
  - "spurious" words: generated from 'nothing' (0:1)
  - One word translated as several words (1:N)

`Canada` 's `program` `has` `been` `implemented`

`Le` `programme` `du` `Canada` `a` `été` `mis` `en` `application`
IBM Model 1 assumption

\[ P(\text{Canada's program has been implemented}) \]

\[ P(\text{Canada's program has been implemented}) \]

\[ = \]

\[ P(\text{Canada's program has been implemented}) \]
IBM Model 1: EM

1. **Initialize** translation parameters randomly (or uniformly).

2. **Expectation**: Compute expected value of $\text{Count}(e, f)$ for all words in training data $\mathcal{O}$, given your current translation parameters, $\theta_k$.

3. **Maximization**: Compute the maximum likelihood estimate of the parameters based on the expected counts, giving improved parameters, $\theta_{k+1}$. 
IBM Model 1: EM

1. Take the product of each conditional word alignment for all alignments and sentence pairs.

2. Normalize by summing over all alignments for each sentence.

3. Add the appropriate normalized counts for each French/English word pair.

(See lecture 5-3)
IBM Model 1: EM

• How does **greedy decoding** work at an abstract level?
  • Consider some of the **transformation functions**.

• How does **phrase-based translation** differ from word-based translation?
  • E.g., we learn alignments given fully observable models in which word alignments are given.
BLEU

• In lecture 6-1, \( ||\text{Ref1}|| = 16, ||\text{Ref2}|| = 17, ||\text{Ref3}|| = 16, \) and \( ||\text{Cn1}|| = 18 \) and \( ||\text{Cn2}|| = 14, \)

\[
brevity_1 = \frac{17}{18} \quad BP_1 = 1
\]

\[
brevity_2 = \frac{16}{14} \quad BP_2 = e^{1-\left(\frac{8}{7}\right)} \approx 0.8669
\]

• Final score of candidate \( C: \)

\[
\text{BLEU} = BP \times (p_1p_2 \ldots p_n)^{1/n}
\]

where
**BLEU example**

- **Reference 1:** I am afraid Dave
- **Reference 2:** I am scared Dave
- **Reference 3:** I have fear David
- **Candidate:** I fear David

- \[ \text{brevity} = \frac{4}{3} \geq 1 \text{ so } BP = e^{1 - \left(\frac{4}{3}\right)} \]

- \[ p_1 = \frac{\sum_{\text{1gram} \in C} \text{Count}_R(1\text{gram})}{\sum_{\text{1gram} \in C} \text{Count}_C(1\text{gram})} = \frac{1+1+1}{1+1+1} = 1 \]

- \[ p_2 = \frac{\sum_{\text{2gram} \in C} \text{Count}_R(2\text{gram})}{\sum_{\text{2gram} \in C} \text{Count}_C(2\text{gram})} = \frac{1}{2} \]

- \[ \text{BLEU} = BP \left( p_1 p_2 \right)^{\frac{1}{2}} = e^{1 - \left(\frac{4}{3}\right)} \left(\frac{1}{2}\right)^{\frac{1}{2}} \approx 0.5067 \]

*Assume \( \text{cap}(n) = 2 \) for all \( n \)-grams*
Automatic speech recognition
What is sound?

- A single **tone** is a sinusoidal function of pressure and time.
  - **Amplitude**: $n$. The degree of the displacement in the air. This is similar to ‘loudness’.
    Often measured in **Decibels (dB)**.
  - **Frequency**: $n$. The number of cycles within a unit of time.
    e.g., **1 Hertz (Hz) = 1 vibration/second**

![Diagram of sound waves showing lower frequency with higher amplitude and higher frequency with lower amplitude.]
Discrete signal representation

**Sampling**: *vbg.* measuring the amplitude of a signal at regular intervals (e.g., 44.1 kHz (CD), 8 kHz (telephone)).

- These amplitudes are initially measured as continuous values at discrete time steps.
Manners of articulation

- **Phoneme**: *n.* a distinctive unit of speech sound.

- **English phonemes can be partitioned into groups, e.g.,**:
  - **Stops/plosives**: complete vocal tract constriction and burst of energy (e.g., ‘**papa**’).
  - **Fricatives**: noisy, with air passing through a tight constriction (e.g., ‘**shift**’).
  - **Nasals**: involve air passing through the nasal cavity (e.g., ‘**mama**’).
  - **Vowels**: open vocal tract, no nasal air.
  - **Glides/liquids**: similar to vowels, but typically with more constriction (e.g., ‘**wall**’).
Windowing and spectra

- Within each window, we extract the amplitudes of the sinusoids that combine to produce the waveform.
- This graph is called the **spectrum**.
Spectrograms

- **Spectrogram**: *n.* a 3D plot of amplitude and frequency over time.
Formants and phonemes

- **Formant**: *n.* A large concentration of energy within a band of frequency (e.g., $F_1$, $F_2$, $F_3$).
The vowel trapezoid

If I asked you about phonemes, I’d probably give you example words.

e.g., iy as in sheet
Prosody

- **Sonorant**: *n.* Any *sustained* phoneme in which the *glottis* is vibrating (i.e., the phoneme is ‘*voiced*’).
  - Includes some consonants (e.g., /w/, /m/, /g/).

- **Prosody**: *n.* the *modification* of speech acoustics in order to convey some *extra-lexical* meaning:
  - **Pitch**: Changing of $F_0$ over time.
  - **Duration**: The length in time of sonorants.
  - **Loudness**: The amount of *energy* produced by the *lungs*. 
Mel-frequency cepstral coefficients

- Mel-frequency cepstral coefficients (MFCCs) are the most popular representation of speech used in ASR.
- They are the spectra of the logarithms of the mel-scaled filtered spectra of the windows of the waveform.

- Based on what we know about human perception of sound and the source-filter model.
Classifying speakers

• Similarly, all of the speech produced by one speaker will cluster differently in MFCC space than speech from another speaker.
• We can decide if a given observation comes from one speaker or another.

\[
P(\text{Observation} | \text{Speaker}_1) > P(\text{Observation} | \text{Speaker}_2)
\]
Continuous distributions

• In the past, we used discrete probability functions.
• Since we are now operating with continuous variables, we need to fit continuous probability functions to a discrete number of observations.

• If we assume the 1-dimensional data in this histogram is Normally distributed, we can fit a continuous Gaussian function simply in terms of the mean $\mu$ and variance $\sigma^2$. 
Mixtures of Gaussians

- Gaussian mixture models (GMMs) are a weighted linear combination of $M$ component Gaussians, $\langle \Gamma_1, \Gamma_2, \ldots, \Gamma_M \rangle$ such that

$$P(\tilde{x}) = \sum_{j=1}^{M} P(\Gamma_j)P(\tilde{x}|\Gamma_j)$$
Continuous HMMs

- Previously we saw **discrete HMMs**: at each state we observed a discrete symbol from a finite set of discrete symbols.
- A **continuous HMM** has observations that are distributed over continuous variables.
  - Observation probabilities, $b_i$, are also continuous.
Levenshtein distance

Allocate matrix $R[n + 1, m + 1]$ \hspace{1em} // where $n$ is the number of reference words
\hspace{1em} // and $m$ is the number of hypothesis words

Initialize $R[0,0] := 0$, and $R[i,j] := \infty$ for all $i = 0$ or $j = 0$

for $i := 1..n$ // #ReferenceWords
  for $j := 1..m$ // #Hypothesis words
    $R(i,j) := \min( R[i-1,j]+1, \hspace{1em} // \text{deletion}$
    $R[i-1,j-1], \hspace{1em} // \text{if the } i^{th} \text{ reference word and}$
    $R[i-1,j-1]+1, \hspace{1em} // \text{if they differ, i.e., substitution}$
    $R[i,j-1]+1 ) \hspace{1em} // \text{insertion}$

Return $100 \times R[n,m]/n$

- See the example in lecture 9-1. Work it out yourself.
Speech synthesis
Speech synthesis

- **Text-to-speech**: *n.* the conversion of electronic text into equivalent, audible speech waveforms.

- Three architectures for performing speech synthesis:
  - Formant synthesis,
  - Concatenative synthesis,
  - Articulatory synthesis.

- How do they differ? What are their (dis)advantages?

- Common components of speech synthesis:
  - **Letter-to-sound rules** and dictionaries,
  - Acoustic prosody modification.
Information retrieval
Information retrieval (IR)

- Given queries in natural language, search for documents or information that answers those queries.
  - Returning documents vs. answering the questions directly.
- Evaluating multiple IR systems using precision and recall.
- The vector space model.
- High-level aspects of singular-value decomposition
The cosine measure

- The **cosine measure** (a.k.a., ‘normalized correlation coefficient’) is

\[
\cos(\mathbf{q}, \mathbf{f}) = \frac{\sum_{i=1}^{n} q_i f_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \sqrt{\sum_{i=1}^{n} f_i^2}}
\]

where \(\mathbf{q}\) and \(\mathbf{f}\) are \(n\)-dimensional vectors for the query and document, respectively.

- **Larger** values of \(\cos(\mathbf{q}, \mathbf{f})\) means **stronger** correlation, so \(\mathbf{q}\) is ‘closer’ to \(\mathbf{f}_1\) than \(\mathbf{f}_2\) iff \(\cos(\mathbf{q}, \mathbf{f}_1) > \cos(\mathbf{q}, \mathbf{f}_2)\).
Precision is not enough

- Each of the following systems has a precision of 50%.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Boogle</th>
<th>Ging</th>
<th>Whoopie</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>3</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>5</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Summarization
Summarization

- Reducing a single document or multiple documents down to their most important or salient elements.
  - **Extractive** summarization vs. **synthetic** summaries.
  - What features are useful in identifying important phrases or sections? What are their properties?
Determining relevance

• The relevance of sentences and phrases within the text can be approximated by:

  • **Position:** The location of the phrase in the document.
  
  • **Cues:** The presence of certain words that indicate relevance (e.g., “crucially”, “in conclusion”).
  
  • **Cohesion:** The distribution of words and their co-occurrences across the document.
ROUGE-2 example

• Candidate: *An egg falls off a wall.*

\[
ROUGE_2 = \frac{\sum_{S \in \{RefSumm\}} \sum_{bigram \in S} \text{Count}_{match}(bigram)}{\sum_{S \in \{RefSumm\}} \sum_{bigram \in S} \text{Count}(bigram)}
\]

\[
ROUGE_2 = \frac{2 + 1 + 0}{8 + 7 + 5} = \frac{3}{20}
\]

Don’t sit on a wall if you’re an egg.

Horses fail to perform surgery upon an egg.

Humpty Dumpty had a great fall.

Ref 1

Ref 2

Ref 3
Miscellaneous classification
Miscellaneous classification

• Walk through and understand the high-level aspects of these models:
  • Neural networks,
  • Support vector machines,
  • Transformation-based learning,
  • $K$ nearest neighbours.

• **Hint**: How do these models differ and how they are similar? What are their strengths and weaknesses? Are there any that are associated with a particular task?
Final thoughts

• This course **barely** scratches the surface of **natural language computing**. Talk to these people:

  - Graeme Hirst
  - Gerald Penn
  - Frank Rudzicz
  - Suzanne Stevenson

• Most of the techniques in this course are applicable **generally**.
  • Hidden Markov models, e.g., are used almost universally, including in finance, biology, medicine, and robotics.
Advice from Mithrandir

• If you don’t study...

YOU! PROBABLY! WON’T! PASS!
Aside – Knowledge

• **Anecdotes** are often useless except as proofs by contradiction.
  • E.g., “I saw Google used as a verb” does not mean that Google is always a verb, just that it is not always a noun.

• **Shallow statistics** are often not enough to be truly meaningful.
  • E.g., “My ASR system is 95% accurate on my test data. Yours is only 90% accurate, idiot.”
    • What if the test data was biased to favor my system?
    • What if we only used a very small amount of data?

• We need a test to see if our statistics actually mean something.

We must be eager to seek evidence that tests our assumptions.
Thank you