Recall our question: how do you tell if two "things" are different?

(Statistical)

What is a language model for? 
- For the purposes of this class:
  - giving a compact representation of what the language is "like".
  - comparing two language sources. (e.g., no country).

A language model $\theta$ is a distribution over all strings $w$ that we wish to consider "legal".

- some distributions aren't made best per word?
  - ex: HMMs (for pronunciation)

A finite non-empty set of states $q_1, q_2, \ldots, q_n$.

- each state $q_i$ has an associated emission distribution $\theta(q_i \rightarrow \epsilon)$, not in $\Lambda$.

- each state $q_i$ has a transition distribution over all other states, $s'$.

- a good counterexample to keep in mind: the Poincaré, over strings of form "jkl" is $\theta$ to be useful, it has no equivalent PCs.

- find $\theta$ for parametric, no equivalent PCs.

(see Booth, Thompson, '73)

- a sequence of states from $\Lambda$ including the empty string.

* A set of sequences of all states from $\Lambda$, including the empty string.

Draw 1st L->N

- Decoding only, e.g., a blink.

$p(s) = \sum$ of $p(s | q_i) p(q_i)$ over $s$, $s \in \{\text{begin}, \text{end}\}$.

$p(s) = \sum$ of $p(s | q_i) p(q_i)$ over $s$, $s \in \{\text{begin}, \text{end}\}$.

$\theta$ should make clear difference between termination and non-termination, and the latter itself being a part of what defines a path.
What does this tell you?
- broken distribution over a potentially uncountable set of sequences to study and define
- by first set of parameters
  (indeed, you have a model), so you're managing the complexity.
- structure of this model gives you insight into what is preferred.

Learning algorithms for HMMs, either from labeled or unlabeled data,
(allow you have to set the # of states beforehand)

- learnable
  - two training data
  - prior
  - need to specify Μ ahead of time.
  - Baum-Welch/EM for max-likelihood estimation (tends to find saddle point or local minimum, be a lot
    - REMEMBER: need constraints, otherwise just sit all params to 00)
  - why are n-gram models so much more common than HMMs?
    - here the data basically has stuck labels!
      - nothing is really "hidden"
    - p(cshd, dog, dog) vs. p(cshd, dog)

- vs. unigram model:

- minor dependence of what is to be next generated

- can build more
  - back into not just fixed state
  - we aren't just stuck to finite state models

- probabilistic context-free grammars (PCFGs) - analogous deference for derivation grammars
  - aka stochastic CFGs (SCFGs)
  - each category type has a distribution over its decompositions, a finite set:

```
\text{ex: } \text{P}\left( \text{S}\text{u}\text{t } \text{N}\text{P}\text{, N}\text{P } \text{S}\text{-qestin} \right)
```

- conditioned
- \text{p}(s) = \text{sum of probs of all trees generating } s
- product of each decomposition used in a tree

```
\text{e.g.: } \text{P}\left( \text{S}\text{u}\text{t } \text{N}\text{P}\text{, N}\text{P } \text{S}\text{-qestin} \right) = \text{P}(\text{sent}) \cdot \text{P}(\text{NP}) \cdot \text{P}(\text{VP}) \cdot \text{P}(\text{they})
```

- [Booth, Thompson '73] conditions under which you get a proper distribution over all strings.

q: does this make sense? Would it make sense to continue on?
- [Booth, Thompson '73] conditions under which you would get a proper distribution over all strings.

[as comments re: PCFG]

-
For this $\Theta = (p)$, that's all we need:

- $P_0$ ("police") = $a$
- $P_0$ ("police police") = $P_0 \cdot P_0 = P_0^2$
- $P_0$ ("police police police") = $[P_0 \cdot P_0 \cdot P_0] + [P_0 \cdot P_0 \cdot P_0^2] = 2P_0^3$
- $\sum_{i=0}^{\infty} P_i(a) = \sum_{i=0}^{\infty} \left( \frac{p}{1-p} \right)^i$ [number of brackets of length $i$ + square] $\sum_{i=0}^{\infty} \left( \frac{p}{1-p} \right)^i$
- (presumably the same as geometric sum?)
- $\sum_{i=0}^{\infty} \left( \frac{p}{1-p} \right)^i = \frac{1}{1 - \left( \frac{p}{1-p} \right)}$

(apparently converges to $\min(1, \frac{q}{p})$)

If $p > \frac{1}{2}$, this will not be proper (too much probability on the "expansion" $P_0$ as opposed to the "generation of words".)

Ex: content model - [Bengo's, 2004] - codec available (also Alexander's)

- (learn) how topics relate to each other w/in docs: potentially relevant to project: (probably have fixed set of topics)

(cks. st.st)

Talked about LLM or just top-most frequent words? [stylistic]

What do you do w/ the "other" words to make then not dominate?

2nd attempt, using Chi's Gamma approach:

- let $X_d$ = prob of all trees w/ depth $\leq d$ (i.e., fringes contain only words)

$$\text{let } X_d = \text{prob of all trees w/ depth } \leq d \text{ (i.e., fringes contain only words)}$$

$$X_{d+1} = p \cdot X_d \cdot X_d$$

If we assume convergence, then we desire for $X = a + pX_0^2$, or $0 = px^2 - x + a$. Solutions are $x = \frac{p}{2}$ or $x = 1$.