What are n-grams good for?

Language ID

(see transparencies)

- Useful for search engines, indexers, etc.
- Useful for text-to-speech (how to pronounce the name "Jan Lukasiewicz"?)

How to Build a Language IDer?

- Histogram of letters?
- Histogram of bigrams?
- Compare to histograms from known language text
- But how to aggregate the evidence?
  - Could steal techniques from information retrieval
  - Treat every language as a huge document and input text as a query

- **Nicer solution:**

  How likely is this text to be generated randomly?

Text Categorization

- Automatic Yahoo classification, etc.
- Similar to language ID …
  - Topic 1 sample: In the beginning God created …
  - Topic 2 sample: The history of all hitherto existing society is the history of class struggles. …
- Input text: Matt's Communist Homepage. Capitalism is unfair and has been ruining the lives of millions of people around the world. The profits from the workers' labor …
  - Input text: And they have beat their swords to ploughshares, And their spears to pruning-hooks. Nation doth not lift up sword unto nation, neither do they learn war any more. …

Topic Segmentation

- Break big document or media stream into indexable chunks
- From NPR's *All Things Considered*:

  The U. N. says its observers will stay in Liberia only as long as West African peacekeepers do, but West African states are threatening to pull out of the force unless Liberia's militia leaders stop violating last year's peace accord after 7 weeks of chaos in the capital, Monrovia … Human rights groups cite peace troops as among those smuggling the arms. I'm Jennifer Ludden, reporting. Whitewater prosecution witness David Hale began serving a 28-month prison sentence today. The Arkansas judge and banker pleaded guilty two years ago to defrauding the Small Business Administration. Hale was the main witness in the Whitewater-related trial that led to the convictions …

Contextual Spelling Correction

- Which is most probable?
  - … I think they're okay …
  - … I think there okay …
  - … I think their okay …

- Which is most probable?
  - … by the way, are they're likely to …
  - … by the way, are there likely to …
  - … by the way, are their likely to …

Speech Recognition

- How do you wreck a nice beach?
- How do you recognize speech?

- Put the file in the folder
- Put the file and the folder

Language generation

- Choose randomly among outputs:
  - Visitant which came into the place where it will be Japanese has admired that there was Mount Fuji.
- Top 10 outputs according to bigram probabilities:
  - Visitors who came in Japan admire Mount Fuji.
  - Visitors who came in Japan admires Mount Fuji.
  - Visitors who arrived in Japan admire Mount Fuji.
  - Visitors who arrived in Japan admires Mount Fuji.
  - Visitors who came to Japan admire Mount Fuji.
  - A visitor who came in Japan admire Mount Fuji.
  - The visitor who came in Japan admire Mount Fuji.
  - Visitors who came in Japan admire Mount Fuji.
  - The visitor who came in Japan admires Mount Fuji.
  - Mount Fuji is admired by a visitor who came in Japan.

Machine Translation

	good English? (n-gram)	good match to French?
Jon appeared in TV.		✓
Appeared on Jon TV.		
In Jon appeared TV.		✓
Jon is happy today.	✓	
Jon appeared on TV.	✓	✓
TV appeared on Jon.	✓	
TV in Jon appeared.		
Jon was not happy.	✓	