Project

EECS 395/495 Probabilistic Graphical Models
Fall 2014

Statistical Language Modeling

- Statistical language models assign **probabilities** to **sequences of words**

$$P(\text{“the dog barked”}) = 4.203 * 10^{-9}$$

- Applications
  - Speech Recognition
  - Machine Translation
  - Spelling Correction
  - Information Extraction

Information Extraction

- IE: Text ➔ machine-understandable data

`Paris, the capital of France, …`

➔

(`Paris`, `France`) $\in$ `CapitalOf`, p=0.85

- Applied to Web: better search engines, semantic Web, step toward human-level AI

IE Automatically?

Intractable to get human labels for every concept expressed on the Web

Idea: extract from **semantically tractable** sentences

…**Edison invented the light bulb**…

(**Edison, light bulb**) ∈ **Invented**

x V y => (***x***, ***y***) ∈ ***V***

…**Bloomberg**, **mayor** of **New York City**…

⇒(**Bloomberg, New York City**) ∈ **Mayor**

x, ***C*** of ***y*** => (***x***, ***y***) ∈ ***C***

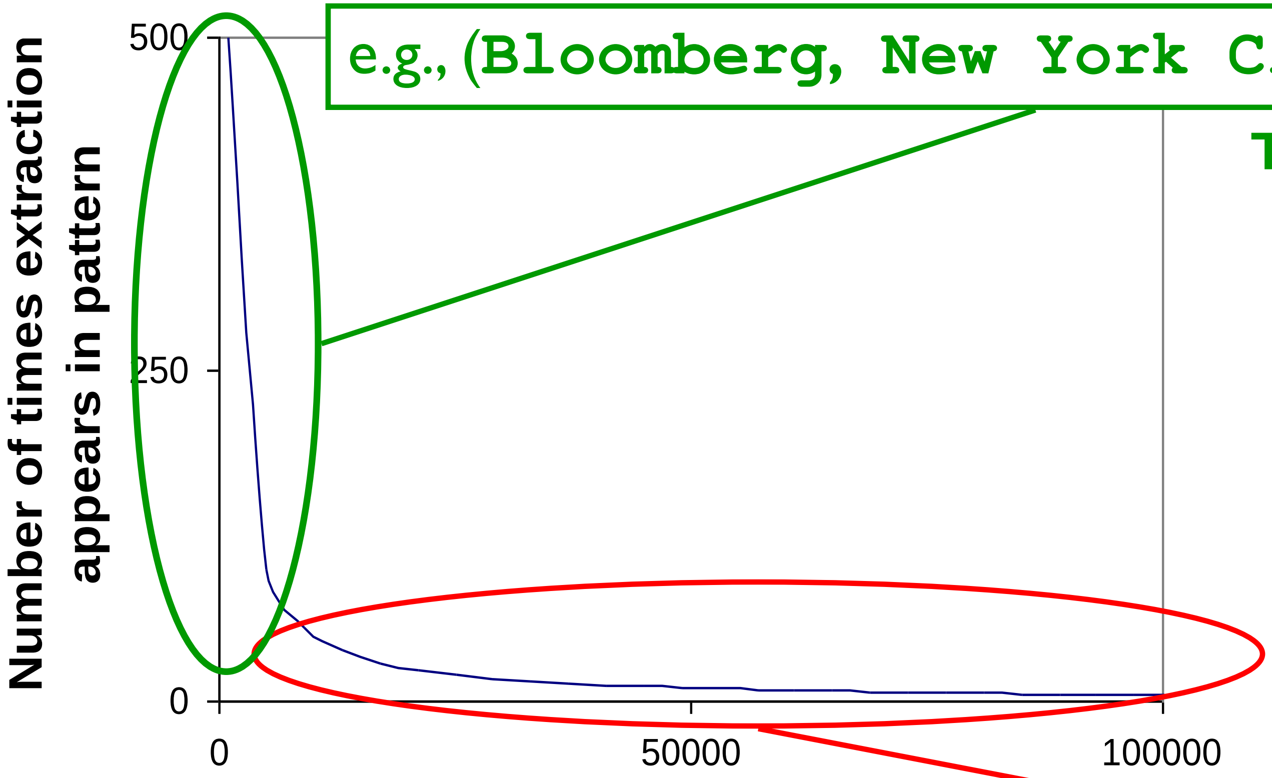
But…

Extraction patterns make errors:

“Erik Jonsson, CEO of **Texas Instruments**, **mayor** of **Dallas** from 1964-1971, and…”

- Empirical fact:
  - Extractions you see over and over tend to be correct
  - The problem is the “long tail”

Challenge: the "long tail"



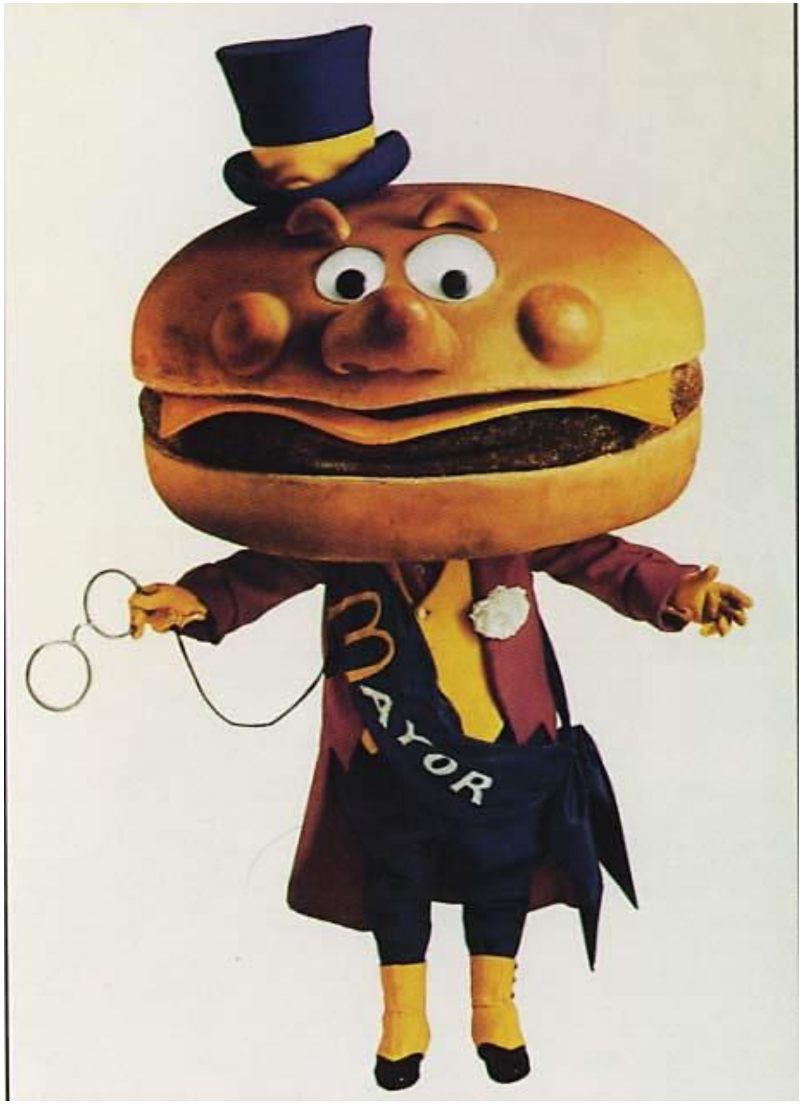
e.g., (**Bloomberg, New York City**)

Tend to be correct

e.g., (**Dave Shaver, Pickerington**)
(**Ronald McDonald, McDonaldland**)

A *mixture* of correct and incorrect

Mayor McCheese



Assessing Sparse Extractions

Idea:

Use statistical language models to determine which sparse extractions are more likely to be correct

Project

- Work in teams of 2-4
  - E-mail me w/ team names and members
- Submit distributions over words for blanks in sentences (demo)
- Do whatever you want, but use probabilistic graphical models
  - We'll discuss a few candidate ideas in class
- Record what works, what doesn't
- Presentations Dec 2, 4 (last week of class)
  - 8 mins + 4 mins Q/A
- Final Report (~2 pages of text + figures/tables)

The Distributional Hypothesis

- *Terms in the same class tend to appear in similar contexts*.

Context	Hits with Chicago	Hits with Twisp
“cities including ___”	42,000	1
“___ and other cities”	37,900	0
“___ hotels”	2,000,000	1,670
“mayor of ___”	657,000	82