Artificial Intelligence

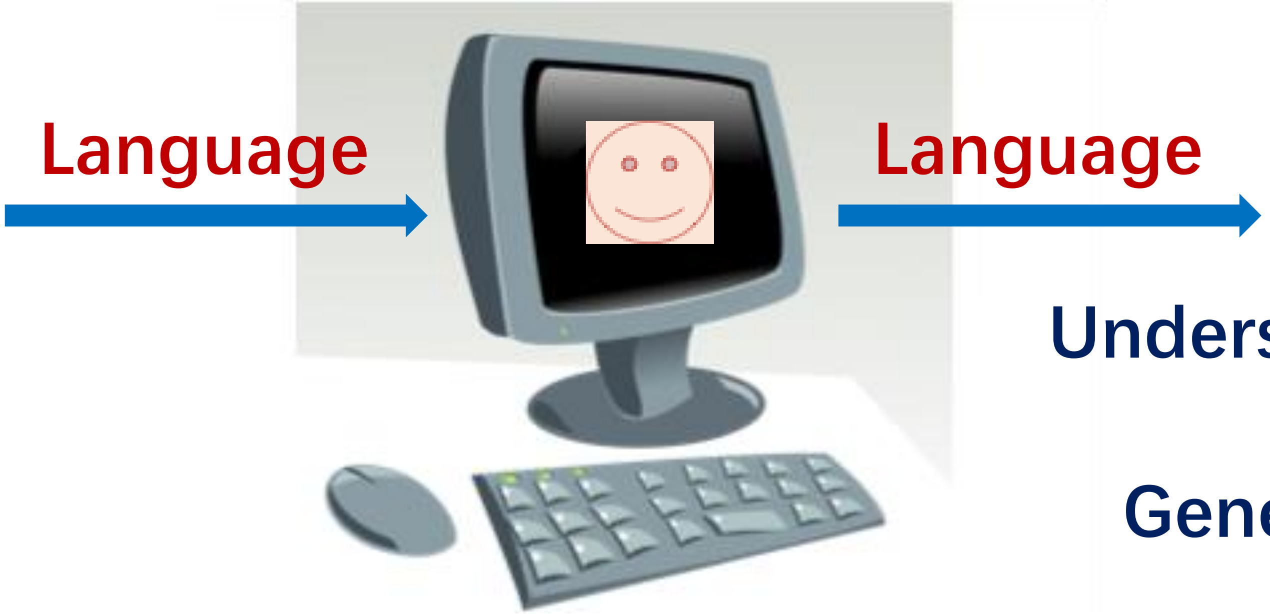
Lecture 15: Natural Language Processing

Credit: Ansaf Salleb-Aouissi, and "Artificial Intelligence: A Modern Approach", Stuart Russell and Peter Norvig, and "The Elements of Statistical Learning", Trevor Hastie, Robert Tibshirani, and Jerome Friedman, and "Machine Learning", Tom Mitchell.

What is NLP

What is Natural Language Processing?

NLP is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human languages.



Understanding language

+

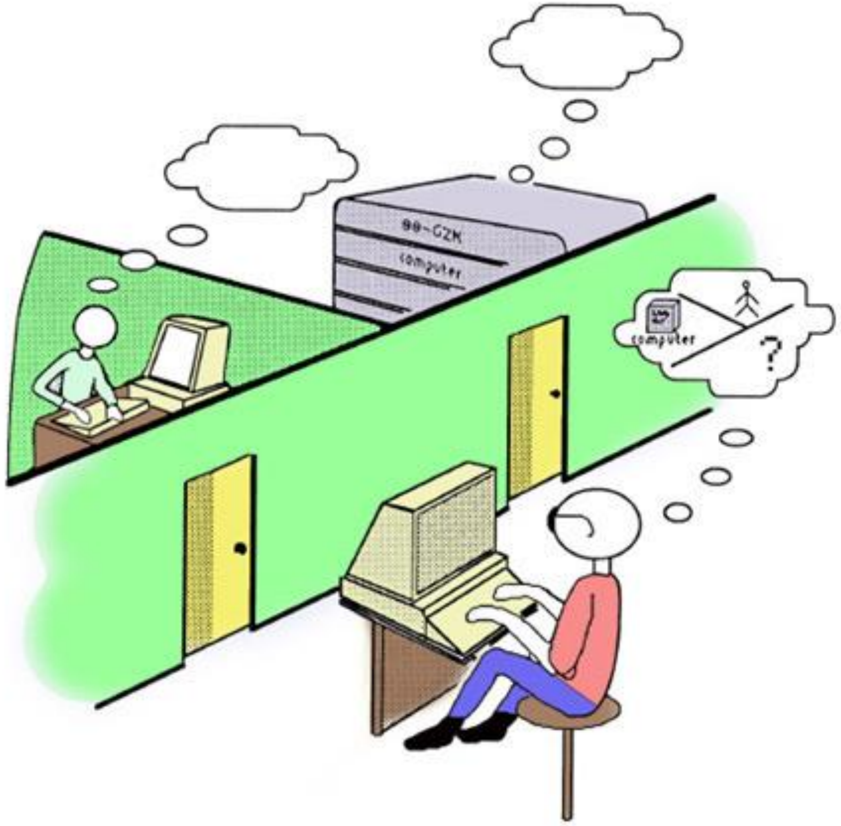
Generating language

NLP

- Natural Language Processing (**NLP**) is an active and attractive field
- Most of our activities online are text-based
- Most of the data available today is text: e-mails, blogs, news, search results, reviews, social media, medical reports, course content, etc.
- Leverage the large and valuable amounts of text available (estimated in hundreds of thousands of perabytes)
- Why NLP? Communicating with computers using natural language has always been a dream…

Acting humanly

- **Turing test (Alan Turing 1950):** A computer passes the test of intelligence, if it can fool a human interrogator.



NLP applications

- **Jeopardy! (2011):** Ken Jennings vs. IBM Watson

Natural language understanding and information extraction!

NLP applications

Speech recognition

- Virtual assistants: Siri (Apple), Echo (Amazon), Google Now, Cortana (Microsoft).
- Leverage deep neural networks to handle **speech recognition** and **natural language understanding.**

NLP applications

Machine translation

- Historical motivation: translate Russian text to English.
- First systems using **mechanical translation (or one-to-one correspondence)** failed!
- "Out of sight, out of mind" → "Invisible, imbecile".

NLP applications

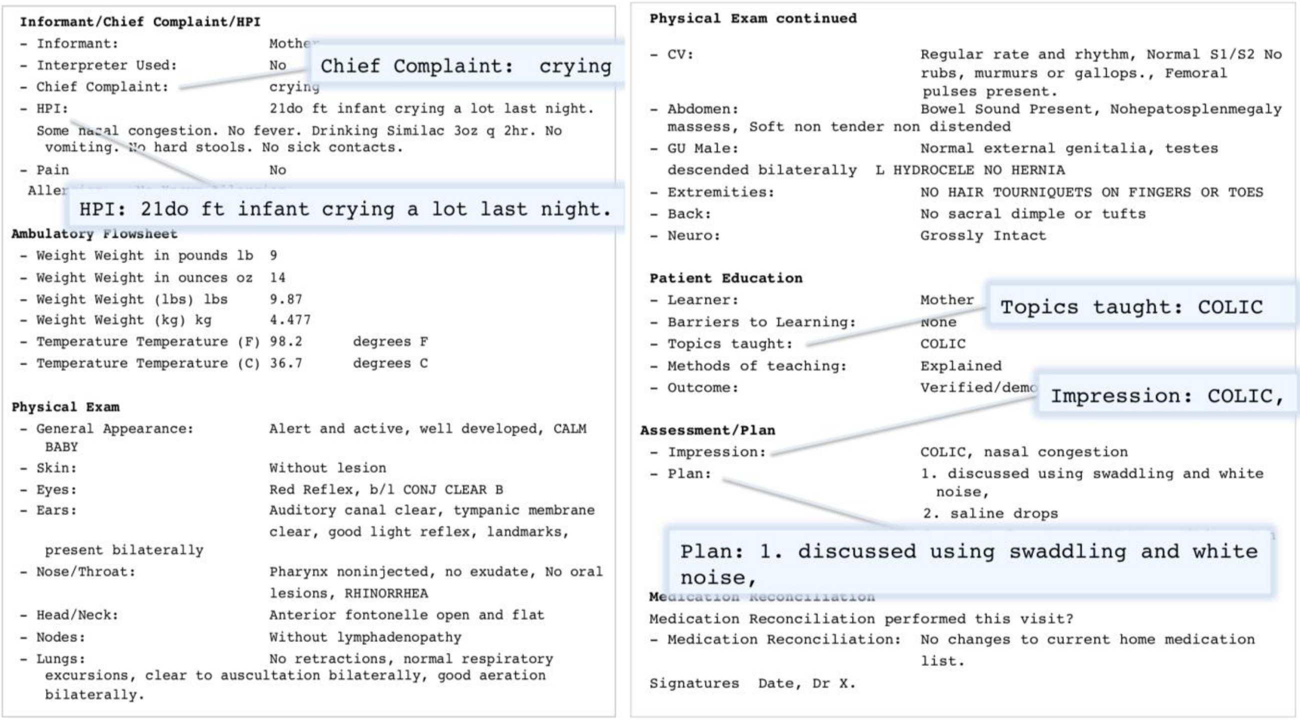
- **Machine translation** has gone through ups and downs.
- Today, Statistical Machine Translation leverages a vast amounts of available **translated corpuses**.
- While there is room for improvement, machine translation has made significant progress.

Google Translate: 100+ languages

NLP applications

Information Extraction

Information extraction: automatically extracting structured information from unstructured or semi-structured text.



NLP applications

Text Summarization

Columbia Newsblaster
Summarizing all the news on the Web

Tuesday, January 5, 2016
Articles from 01/02/2016 to 01/05/2016
Last update: 2:43 PM EST

Search for:
Offline summar Go

U.S.
World
Finance
Sci/Tech
Sports

View Today's Images
View Archive
About Newsblaster
About today's run
Newsblaster in Press
Academic Papers

Article Sources:
abcnews.go.com (54 articles)
latimes.com (16 articles)
haaretz.com (10 articles)
foxnews.com (10 articles)
cbsnews.com (6 articles)
baltimoresun.com (5 articles)
usatoday.com (2 articles)
bbc.co.uk (2 articles)

'GMA Ultimate Tailgate Challenge': Rob's Sizzling Sausage and Peppers Video (U.S., 49 articles)
Now Playing: John Stamos Talks New Comedy Series Grandfathered Now Playing: Brie Larson Stars in Critically Acclaimed Room Now Playing: John Krasinski Hits the Big Screen in 13 Hours: The Secret Soldiers of Benghazi Now Playing: GMA Ultimate Tailgate Challenge: Rob's Sizzling Sausage and Peppers. Now Playing: Does Fedex Extra Coverage Protect Against Money Lost From Late Packages? Now Playing: What Are the Best Ways to Get Rid of Those Unwanted Gifts? Now Playing: FedEx Employees Play Catch-up to Deliver Late Holiday Packages Now Playing: Could the Oregon Militia Standoff Turn Violent? Now Playing: The Hunt for the New Jihadi John Now Playing: Camille Cosby Seeks to Delay Deposition in Husband's Sexual Assault Case. Now Playing: LG Announces New TV Now Playing: More Than 50 Homes Damaged in Oklahoma City House Explosion. Now Playing: Will 2016 Bring Self-Driving Cars and Room Service Robots? Now Playing: 46 Years of Friendship on Facebook? Now Playing: Apple iPhone 6S Wins Top Smartphone Award. Now Playing: Stuntman Describes Death-Defying Wingsuit Scene in Point Break Now Playing: Go Motocrossing With Stuntmen From Point Break Now Playing: One on One With Point Break Star Edgar Ramirez.

Top News

Iran-Saudi Arabia row: Kuwait recalls ambassador from Tehran (World, 7 articles) [UPDATE]
Oil prices jumped on the first trading day of 2016 as Middle East tension outweighed a selloff in financial markets around the world. The conflict between Iran and Saudi Arabia has simmered for months, with the wars in Yemen and Syria playing out as proxy fights between the two rivals. The execution last weekend of Sheikh Nimr al-Nimr, a Shiite cleric and opposition figure in Saudi Arabia, has heightened the Saudi-Iran regional rivalry, threatening to derail already-shaky peace efforts over the wars in Syria and Yemen.

Mitch Kupchak says Lakers may retire both '8' and '24" for Kobe Bryant (Sports, 8 articles) [UPDATE]
With their Sunday night win over the Phoenix Suns, the Lakers are on a three-game win streak with eight victories in 35 tries, slotting the team last place in the Western Conference. Even though the Golden State Warriors are coming to town Tuesday for a true reality check, there was more interest Monday in the prickly relationship between Lakers Coach Byron Scott and Julius Randle . The 21-year-old reserve power forward did not like being mentioned by Scott for playing poor defense after the Lakers' 97-77 victory Sunday against the Phoenix Suns.

Like old times, Bill Clinton joins the campaign trail in New Hampshire (U.S., 7 articles) [UPDATE]
President Bill Clinton s political muscle memory took him down a well-worn path Monday in New Hampshire: the rally in Nashua, the lunchtime mingle in Manchester and the afternoon town hall in Exeter. Vermont Sen. Bernie Sanders will pledge to break up the country's largest financial institutions within the first year of his administration should he win the White House next November. We begin this evening with the all out sprint to Iowa one month now until the first votes and New Hampshire of course shortly thereafter.

Israel News (Science/Technology, 6 articles)
BREAKING NEWS 7:38 PM 4:50 PM 4:49 PM 4:49 PM 4:49 PM 4:34 PM 4:32 PM 3:05 PM 2:44 PM 2:43 PM 2:42 PM 2:26 PM 2:23 PM 2:18 PM 2:12 PM More Breaking News

Finance

- U.S. media stocks dip following China market rout (5 articles) [UPDATE]

Sports

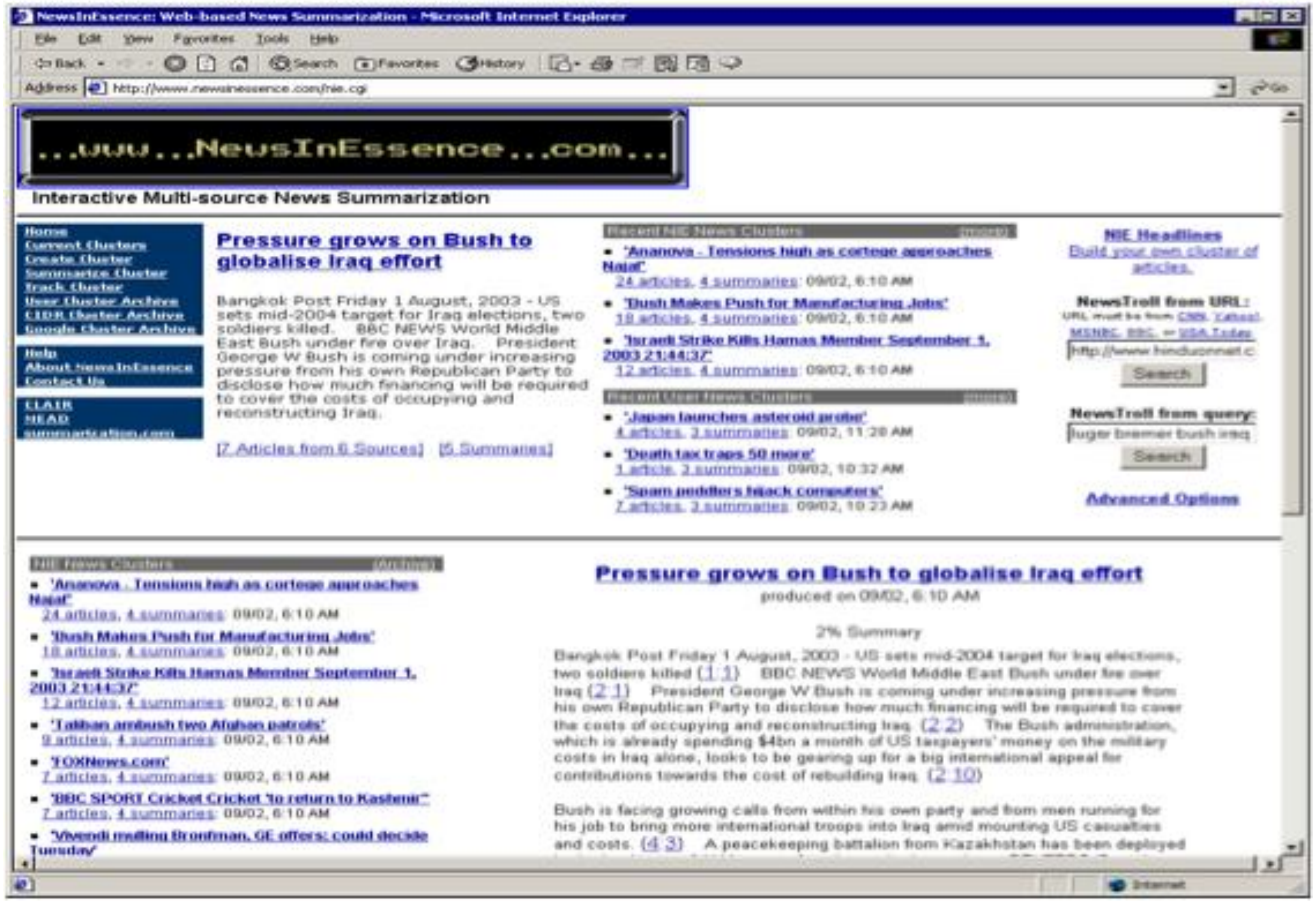
- Bieksa, Kesler score on power play; Ducks beat Jets (4 articles) [UPDATE]
- LAS VEGAS, NV - JANUARY 02: (R-L) Robbie Lawler exchanges punches with Carlos Condit in their welterweight championship fight during the UFC 195 event inside MGM Grand Garden Arena on January 2, 2016 in Las Vegas, Nevada. (Photo by Brandon Magnus/Zuffa LLC/Zuffa LLC via Getty Images) (4 articles) [UPDATE]

Science/Technology

blaster@cs.columbia.edu

NLP applications

Text Summarization



NLP applications

Dialog systems

e.g., automated online assistants.

Caller: I need to check my account status.
System: What is your name?
User: Goodhanilobees
System: I didn' t get that. Please spell your name
User: G.o.o.d.h.a.n.i.l.o.b.e.e.s.
System: I still didn' t get that. Please spell your name again
Caller: An agent PLEASE! NOW!
System: All our agents are assisting other customers... but I am an agent too! an intelligent agent...

NLP applications

Sentiment Analysis

☆☆☆☆☆ Fantastic... truly a wonderful family movie

☆☆☆ I have a mixed feeling about this movie.

☆☆☆ Well it is fun for sure but definitely not appropriate for kids 10 and below

☆☆☆☆☆ My kids loved it!!

☆☆☆☆☆ The movie is very funny and entertaining. Big A+

☆ I got so boooored...

☆☆ Disappointed. They showed all fun details in the trailer

☆☆☆ Cute but not for adults

NLP & AI

NLP is one of the hardest problems in Artificial Intelligence.

Human language is so complex!

NLP

1. **Ambiguity:**
   "At last, a computer that understands you like your mother."
   1985 McDonnell-Douglas ad.
2. **Anaphora:** He bought a brand new car and drove it home.
3. **Metonymy:** She learned how to play Mozart at a very young age.
4. **Metaphor:** He is a walking dictionary! His room is a zoo.
5. **Vagueness, discourse structure, auto correction, etc.**

Text Classification

Learning to classify text. Why?

- Learn which news articles are of interest
- Learn to classify web pages by topic
- Classify Spam from non Spam emails
- Naive Bayes is among most effective algorithms
- What attributes shall we use to represent text documents?

Setting

- A training data (x_i, y_i), x_i is a feature vector and y_i is a discrete label. d features, and n examples.
- Example: consider document classification.
- A new example with feature values $x_{new} = (a_1, a_2, \ldots, a_d)$.
- We want to predict the label y_{new} of the new example.

$$y_{new} = \operatorname*{arg\,max}_{y \in \mathbb{Y}} \; p(y|a_1, a_2, \cdots, a_d)$$

Naive Bayes Classifier

Use simplifying assumption:

$$p(a_1, a_2, \cdots, a_d | y) = \prod_j p(a_j | y)$$

Naive Bayes Classifier:

$$y_{new} = \operatorname*{arg\,max}_{y \in \mathbb{Y}} \; p(y) \prod_j p(a_j | y)$$

Algorithm

Learning: Based on the frequency counts in the dataset:

1. Estimate all $p(y)$, $\forall y \in Y$.
2. Estimate all $p(a_j|y)$, $\forall y \in Y$, $\forall a_j$.

Classification: For a new example, use:

$$y_{new} = \text{argmax}_{y \in \mathbb{Y}} p(y) \prod_j p(a_j|y)$$

Note: No model per se or hyperplane, just count the frequencies of various data combinations within the training examples.

Estimating probabilities

m-estimate of the probability: $$p(a_j|y) = \frac{n_c + m * p}{n_y + m}$$

where:

n_y: total number of examples for which the class is y.

n_c: total number of examples for which the class is y and feature x_j = a_j.

m: called equivalent sample size

Intuition:

Augment the sample size by m virtual examples, distributed according to prior p (prior estimate of each value).

If prior is unknown, assume uniform prior: if a feature has k values, we can set p=1/k.

Text Classification

- Given a document (corpus), define an attribute for each word position in the document.
- The value of the attribute is the English word in that position.
- To reduce the number of probabilities that needs to be estimated, besides NB independence assumption, we assume that: The probability of a given word w_k occurrence is independent of the word position within the text. That is:

$$p(x_1 = w_k|c_j), p(x_2 = w_k|c_j), \cdots$$

estimated by:

$$p(w_k|c_j)$$

Text Classification

- m-estimate of the probabilities:

$$p(w_k|c_j) = \frac{n_k + 1}{n_j + |Vocabulary|}$$

  where:

  n_j: total #word positions in all training examples of class c_j.

  n_k: number of times the word w_k is found in among these n_j word positions.

- The following function/algorithms learns the probabilities $P(w_k|c_j)$ describing the probability that a randomly drawn word from a document with class c_j is the English word w_k. It also learn the class priors $P(c_j)$.

Learn_Naive_Bayes_texte(Examples, C)

Input: Examples is a set of document with classes. C is the set of classes.

1. Collect all words, punctuations and tokens occuring in the Examples. Let the pertinent vocabulary be V.

2. Calculate $P(c_j)$ and $P(w_k/c_j)$.

   - For each class c_j in C
     - $docs_j \leftarrow$ the subset of documents from Examples for which the label=c_j
     - $P(c_j) \leftarrow \frac{|docs_j|}{|Examples|}$
     - $text_j \leftarrow$ a single document concatenation of all documents in doc_j
     - $n_j \leftarrow$ total number of distinct word positions in $text_j$
     - for each word w_k in V
       * $n_k \leftarrow$ number of times word w_k appears in $text_j$
       * $P(w_k/c_j) \leftarrow \frac{n_k+1}{n_j+|V|}$

Output: all $P(c_j)$ and $P(w_k/c_j)$.

Classify_Naive_Bayes_text(Doc)

Return the estimated label for the document Doc. a_i denotes the word found in the ith position within Doc.

- positions $\leftarrow$ all word positions in Doc that contain token found in V.
- Return c_{Doc} where:

$$c_{Doc} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in positions} P(a_i/c_j)$$

Example

Classification of Radio and TV sentences.

TV:

1. TV programs are not interesting – TV is annoying.
2. Kids like TV.
3. We receive TV by radio waves.

Radio:

1. It is interesting to listen to the radio.
2. On the waves, kids programs are rare.
3. The kids listen to the radio; it is rare!

Vocabulary:

V = {TV, program, interesting, kids, radio, wave, listen, rare}

Example

$p(C_{Tv}) = 3/6 = 0.5 \qquad p(C_{Radio}) = 3/6 = 0.5$

$n_{TV} = 9 \qquad n_{Radio} = 11$

$w \in \mathcal{V}$	Class "TV"			Class "Radio"		
	n_{TV}	n_w	$p(w\|C_{TV})$	n_{Radio}	n_w	$p(w\|C_{radio})$
TV	9	4	(4+1)/(9+8)	11	0	1/(11+8)
program	9	1	(1+1)/(9+8)	11	1	2/(11+8)
interesting	9	1	(1+1)/(9+8)	11	1	2/(11+8)
kids	9	1	(1+1)/(9+8)	11	2	3/(11+8)
radio	9	1	(1+1)/(9+8)	11	2	3/(11+8)
wave	9	1	(1+1)/(9+8)	11	1	2/(11+8)
listen	9	0	(0+1)/(9+8)	11	2	3/(11+8)
rare	9	0	(0+1)/(9+8)	11	2	3/(11+8)

Language Models

- We just saw that language is complex, there is no single meaning, we disagree on the grammar and there is not set of definitive sentences
- Instead of talking of one single meaning of a sentence, we talk of **probability distribution over meaning**
- A language model is an approximation of language
- Aim: Model natural language

Language Models

"Did you call your …"

- How can we guess or predict the next word?
- Possible words to follow: **mother**, **doctor**, **child**, …
- Unlikely words to follow: **dinosaur**, **oven**, …
- Estimate

$$P(w|Did\ you\ call\ your\ldots)$$

for any w.

Language Models

- **Build a probabilistic language model that assigns a:**
  - probability to each next possible word: **predict** the next word

$$P(mother|Did\ you\ call\ your\ldots)$$

$$P(dinosaur|Did\ you\ call\ your\ldots)$$

$$P(doctor|Did\ you\ call\ your\ldots)$$

  - probability to a complete sentence (sequence of words): **predict** the probability to see this sentence in a text

$$P(Open\ your\ book\ on\ page\ six)$$

$$P(book\ open\ ten\ your\ on\ page)$$

Language Models

Language models are crucial in many NLP applications:

- **Spell correction**

  "Once upon a time" versus "Ounce upon a time"

- **Statistical machine translation**

  "Out of sight, out of mind" translation to either (1) "Invisible, imbecile" or (2) "Hors de vue, hors de l' esprit".

- **Seek information** (text classification, information retrieval, information extraction).
- **Speech recognition**
- **Language identification**

Language Models

N-gram models

- Estimate P(page|open your book on) using frequencies in a large corpus:

$$P(\texttt{page}|\texttt{open your book on}) = \frac{\texttt{count(open your book on page)}}{\texttt{count(open your book on)}}$$

- Estimate P(open your book on page) using frequencies in a large corpus:

$$P(\texttt{open your book on page}) = \frac{\texttt{count(open your book on page)}}{\texttt{count(sentences of 5 words)}}$$

- The corpus has to be very very large!
- Poor model. Will be zero for a sentence that does not appear in the corpus.

Language Models

N-gram models

- **Problem:** How to estimate the joint probability?

$$P(w_1, w_2, \ , \cdots, w_n)$$

- **Solution:** decompose the joint probability using **chain rule of probability**

$$P(w_1, \cdots, w_n) = p(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_1 \cdots w_{n-1})$$

$$P(w_1, \cdots, w_n) = \prod_{k=1}^{n} P(w_k|w_1 \cdots w_{k-1})$$

Language Models

N-gram models

- **Idea:** Instead of using the whole chain, approximate using the last words.
- Bigram model: uses the **Markov assumption**

$$P(w_n|w_{n-1})$$

  to approximate

$$P(w_n|w_1 \ldots w_{n-1})$$

  e.g., $P(page|on)$.
- Trigram model: look two words in the past.
- N-gram model: look into $N - 1$ words in the past.

Language Models

N-gram models

- N-gram:

$$P(w_n|w_1 \cdots w_{n-1}) \approx P(w_n|w_{n-N+1} \cdots w_{n-1})$$

- Bigram:

$$P(w_1, \cdots, w_n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

- Use **Maximum Likelihood Estimate (MLE):**

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}w_n)}{\sum_w count(w_{n-1}w)}$$

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}w_n)}{count(w_{n-1})}$$

Language Models

N-gram models

- Use Maximum Likelihood Estimate (MLE) for N-gram:

$$P(w_n|w_{n-N+1}\cdots w_{n-1}) = \frac{count(w_{n-N+1}\cdots w_{n-1}w_n)}{count(w_{n-N+1}\cdots w_{n-1})}$$

- Bigrams capture syntactic dependencies such as a **noun** comes after **eat**, and that a **verb** comes after **to** etc.
- In practice, using 3-grams and 4-grams are common. We also use log probability to get larger numbers instead of probabilities.

Language Models

Example of bigrams

- Bigram probabilities
  1. $*$ I love cheese STOP
  2. $*$ Cheese and crackers are delicious STOP
  3. $*$ I prefer swiss cheese STOP

$$P(\texttt{I}|*) = \frac{2}{3}$$
$$P(\texttt{Cheese}|*) = \frac{1}{3}$$
$$P(\texttt{STOP}|\texttt{Cheese}) = \frac{2}{3}$$
$$P(\texttt{prefer}|\texttt{I}) = \frac{1}{2}$$

- Probability of a sentence can be obtained by multiplying the the bigram probabilities.

$$P(*\texttt{I eat cheese STOP}) = P(\texttt{I}|*)P(\texttt{eat}|\texttt{I})P(\texttt{cheese}|\texttt{eat})P(\texttt{STOP}|\texttt{cheese})$$

Language Models

Evaluation

- Use a **training corpus** and **test corpus**
- To compare two language models, calculate the probability of the test corpus with both models. Pick the one with a higher probability
- Use **Perplexity**: Inverse probability of the test corpus normalized by the number of words in the test, *N*.

$$\text{Perplexity}(w_1 w_2 \cdots w_N) = P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}}$$

Language Models

Evaluation

- Perplexity

$$\text{Perplexity}(w_1 w_2 \cdots w_N) = P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}}$$

$$\text{Perplexity}(w_1 w_2 \cdots w_N) = (\prod_{i=1}^{N} P(w_i | w_1 \cdots w_{i-1}))^{-\frac{1}{N}}$$

- For bigrams:

$$\text{Perplexity}(w_1 w_2 \cdots w_N) = (\prod_{i=1}^{N} P(w_i | w_{i-1}))^{-\frac{1}{N}}$$

- The higher the conditional probability, the lower the perplexity.
- Empirically, the more information provided by the N-gram, the lower the perplexity (the word sequence is captured).

Language Models

Smoothing

$$P(\texttt{*I eat cheese STOP}) = P(\texttt{I}|\texttt{*})P(\texttt{eat}|\texttt{I})P(\texttt{cheese}|\texttt{eat})P(\texttt{STOP}|\texttt{cheese})$$

- Some probabilities may be zero!
- We modify the N-gram counts:

$$P(w_j) = \frac{\text{count}(w_j)}{N}$$

$$P_L(w_j) = \frac{\text{count}(w_j) + 1}{N + V}$$

Progress in NLP

Big progress

- **Tagging**

  Text ⇒ Tagged Text

  – Part of Speech tagging

    I(P) shoot(V) the(A) wumpus(N)

  – Name Entity Recognition

    Yesterday(time) I(person) bought five(quantity) books from Amazon (Co.)

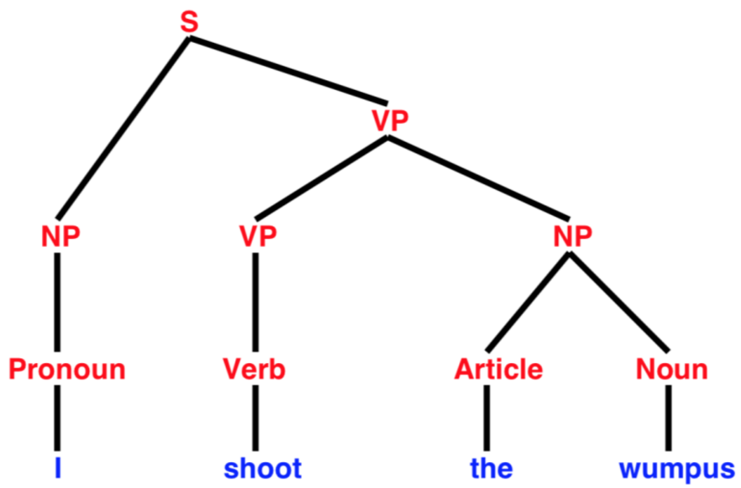
- **Text classification (Spam filtering)**

Progress in NLP

Good progress

- **Parsing**

  Exhibit the grammatical structure of a sentence: Text ⇒ Tree



- **Sentiment analysis**

  √ Fantastic… truly a wonderful family movie

  × I got so boooored…

- **Machine translation**
- **Information extraction**

Progress in NLP

Work in progress

- Summarization

Health Benefits

- Eating a diet rich in vegetables and fruits as part of an overall healthy diet may reduce risk for heart disease, including heart attack and stroke.
- Eating a diet rich in some vegetables and fruits as part of an overall healthy diet may protect against certain types of cancers.
- Diets rich in foods containing fiber, such as some vegetables and fruits, may reduce the risk of heart disease, obesity, and type 2 diabetes.
- Eating vegetables and fruits rich in potassium as part of an overall healthy diet may lower blood pressure, and may also reduce the risk of developing kidney stones and help to decrease bone loss.
- Eating foods such as vegetables that are lower in calories per cup instead of some other higher-calorie food may be useful in helping to lower calorie intake.

Nutrients

- Most vegetables are naturally low in fat and calories. None have cholesterol. (Sauces or seasonings may add fat, calories, or cholesterol.)
- Vegetables are important sources of many nutrients, including potassium, dietary fiber, folate (folic acid), vitamin A, and vitamin C.
- Diets rich in potassium may help to maintain healthy blood pressure. Vegetable sources of potassium include sweet potatoes, white potatoes, white beans, tomato products (paste, sauce, and juice), beet greens, soybeans, lima beans, spinach, lentils, and kidney beans.
- Dietary fiber from vegetables, as part of an overall healthy diet, helps reduce blood cholesterol levels and may lower risk of heart disease. Fiber is important for proper bowel function. It helps reduce constipation and diverticulosis. Fiber-containing foods such as vegetables help provide a feeling of fullness with fewer calories.
- Folate (folic acid) helps the body form red blood cells. Women of childbearing age who may become pregnant should consume adequate folate from foods, and in addition 400 mcg of synthetic folic acid from fortified foods or supplements. This reduces the risk of neural tube defects, spina bifida, and anencephaly during fetal development.
- Vitamin A keeps eyes and skin healthy and helps to protect against infections.
- Vitamin C helps heal cuts and wounds and keeps teeth and gums healthy. Vitamin C aids in iron absorption.

Eating vegetables is healthy.

- Question/Answering
- Dialog Systems: Siri, echo, etc.

To be continued