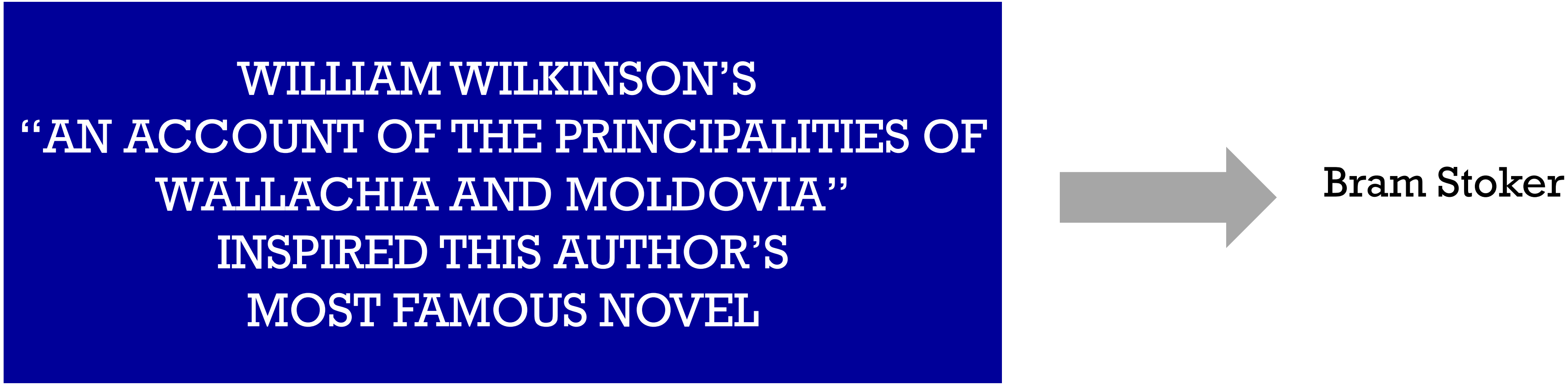
INTRODUCTION TO NLP

What is Natural Language Processing?

QUESTION ANSWERING: IBM'S WATSON

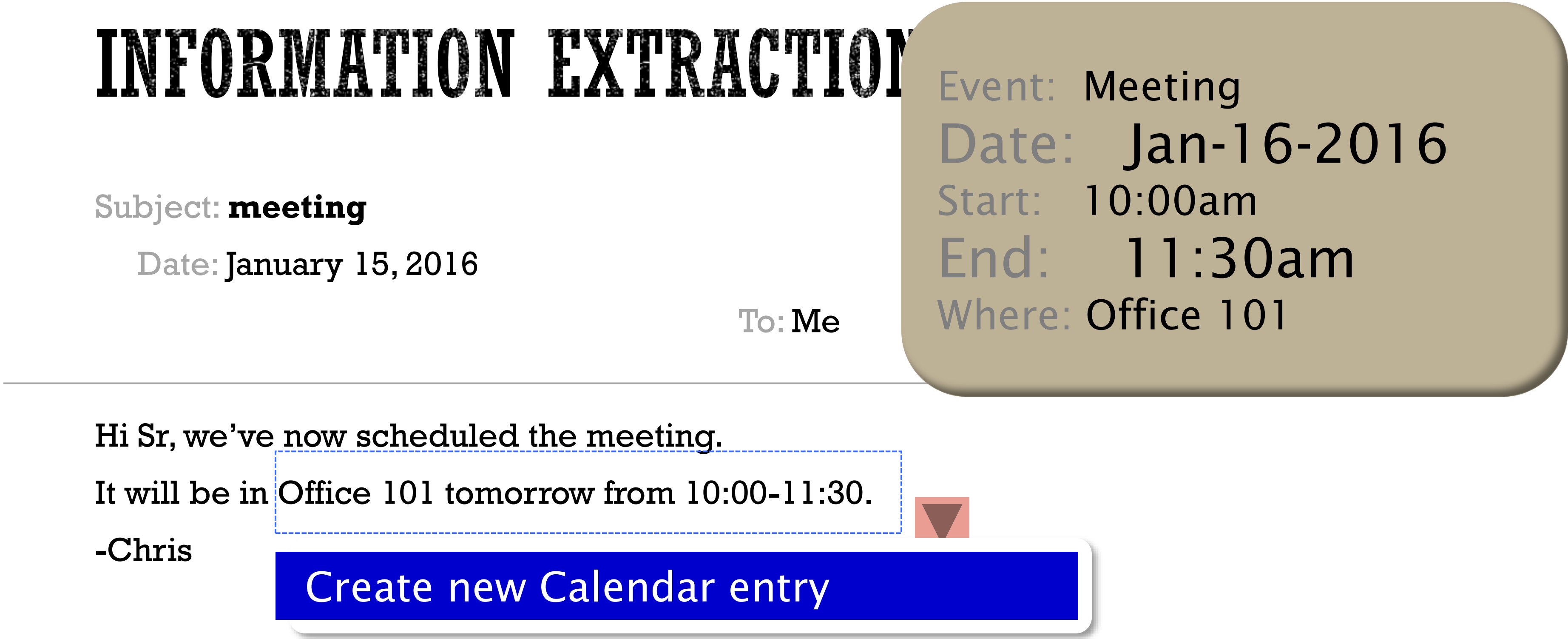
- Won Jeopardy on February 16, 2011!



WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

INFORMATION EXTRACTIO



Subject: **meeting**

Date: January 15, 2016

To: Me

Hi Sr, we've now scheduled the meeting.

It will be in Office 101 tomorrow from 10:00-11:30.

-Chris

Event: Meeting
Date: Jan-16-2016
Start: 10:00am
End: 11:30am
Where: Office 101

INFORMATION EXTRACTION & SENTIMENT ANALYSIS

Attributes:

- zoom
- affordability
- size and weigh
- flash
- ease of use

Size and weight

✓
- nice and compact to carry!

✓
- since the camera is small and light, I won't need to carr professional cameras either!

✗
- the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

MACHINE TRANSLATION

- Fully automatic

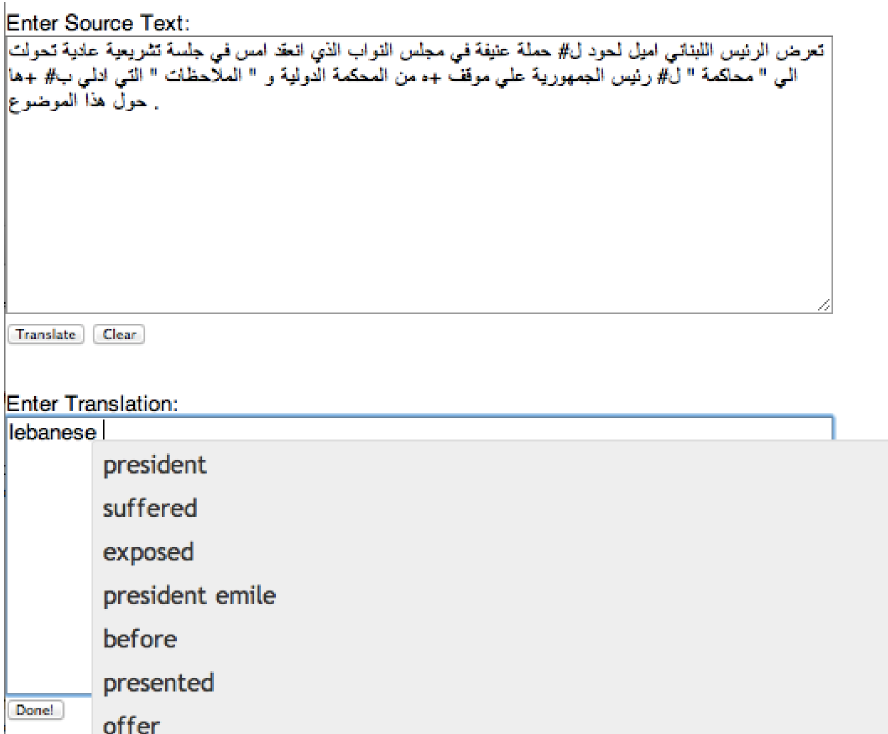
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators



Enter Source Text:

تعرض الرئيس اللبناني اميل لحود ل# حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت
الي " محاكمة " ل# رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي ب# +ها
حول هذا الموضوع .

Enter Translation:

lebanese

LANGUAGE TECHNOLOGY

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON			ORG			LOC
Einstein	met	with	UN	officials	in	Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

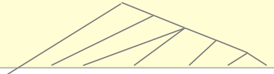
Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕… ⇨

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up
The S&P500 jumped
Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

AMBIGUITY MAKES NLP HARD: "CRASH BLOSSOMS"

100% REAL

Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

AMBIGUITY IS PERVASIVE

New York Times headline (17 May 2000)

Fed raises interest rates

Fed raises interest rates

Fed raises interest rates 0.5%

WHY ELSE IS NATURAL LANGUAGE UNDERSTANDING DIFFICULT?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either ♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

But that's what makes it fun!

MAKING PROGRESS ON THIS PROBLEM…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilistic models built from language data
    - P("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
  - Luckily, rough text features can often do half the job.

INTRODUCTION TO NLP

What is Natural Language Processing?

BASIC TEXT PROCESSING

Word tokenization

TEXT NORMALIZATION

- Every NLP task needs to do text normalization:
  1. Segmenting/tokenizing words in running text
  2. Normalizing word formats
  3. Segmenting sentences in running text

HOW MANY WORDS?

- I do uh main- mainly business data processing
  - Fragments, filled pauses
- Seuss's cat in the hat is different from other cats!
  - **Lemma**: same stem, part of speech, rough word sense
    - cat and cats = same lemma
  - **Wordform**: the full inflected surface form
    - cat and cats = different wordforms

HOW MANY WORDS?

they lay back on the San Francisco grass and looked at the stars and their

- **Type**: an element of the vocabulary.
- **Token**: an instance of that type in running text.
- How many?
  - 15 tokens (or 14)
  - 13 types (or 12) (or 11?)

HOW MANY WORDS?

N = number of tokens

V = vocabulary = set of types

$|V|$ is the size of the vocabulary

Church and Gale (1990): $|V| > O(N^{½})$

	Tokens = N	**Types = |V|**
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

SIMPLE TOKENIZATION IN UNIX

- (Inspired by Ken Church's UNIX for Poets.)
- Given a text file, output the word tokens and their frequencies

```
tr -sc 'A-Za-z' '\n' < shakes.txt
     | sort
     | uniq –c
```

Change all non-alpha to newlines

Sort in alphabetical order

Merge and count each type

```
1945 A
  72 AARON
  19 ABBESS
   5 ABBOT
 ... ...
```

```
25 Aaron
 6 Abate
 1 Abates
 5 Abbess
 6 Abbey
 3 Abbot
.... …
```

THE FIRST STEP: TOKENIZING

```
tr -sc ’A-Za-z’ ’\n’ < shakes.txt | head
```

```
THE
SONNETS
by
William
Shakespeare
From
fairest
creatures
We
...
```

THE SECOND STEP: SORTING

```
tr -sc ’A-Za-z’ ’\n’ < shakes.txt | sort | head
```

```
A
A
A
A
A
A
A
A
A
...
```

ISSUES IN TOKENIZATION

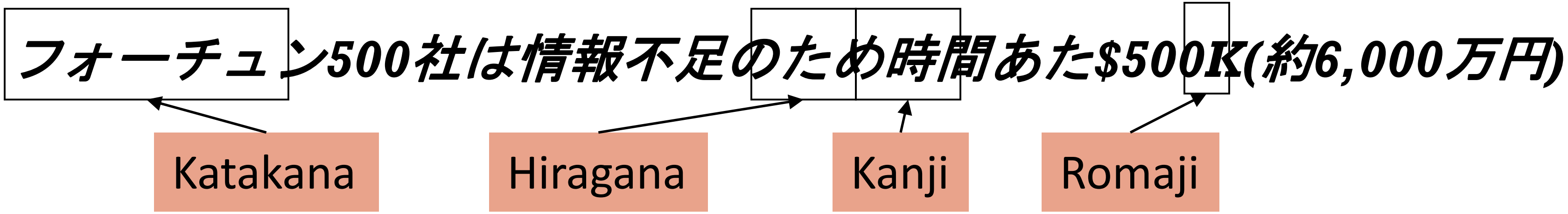
- Finland’s capital → Finland Finlands Finland’s ?
- what’re, I’m, isn’t → What are, I am, is not
- Hewlett-Packard → Hewlett Packard ?
- state-of-the-art → state of the art ?
- Lowercase → lower-case lowercase lower case ?
- San Francisco → one token or two?
- m.p.h., PhD. → ??

TOKENIZATION: LANGUAGE ISSUES

- French
  - ***L'ensemble*** → one token or two?
    - ***L*** ? ***L'*** ? ***Le*** ?
    - Want ***l'ensemble*** to match with ***un ensemble***
- German noun compounds are not segmented
  - ***Lebensversicherungsgesellschaftsangestellter***
  - 'life insurance company employee'
  - German information retrieval needs **compound splitter**

TOKENIZATION: LANGUAGE ISSUES

- Chinese and Japanese no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats



フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

Katakana

Hiragana

Kanji

Romaji

End-user can express query entirely in hiragana!

WORD TOKENIZATION IN CHINESE

- Also called **Word Segmentation**
- Chinese words are composed of characters
  - Characters are generally 1 syllable and 1 morpheme.
  - Average word is 2.4 characters long.
- Standard baseline segmentation algorithm:
  - Maximum Matching (also called Greedy)

BASIC TEXT PROCESSING

Word tokenization

BASIC TEXT PROCESSING

Word Normalization and Stemming

NORMALIZATION

- Need to "normalize" terms
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match ***U.S.A.*** and ***USA***
- We implicitly define equivalence classes of terms
  - e.g., deleting periods in a term
- Alternative: asymmetric expansion:
  - Enter: ***window*** Search: ***window, windows***
  - Enter: ***windows*** Search: ***Windows, windows, window***
  - Enter: ***Windows*** Search: ***Windows***
- Potentially more powerful, but less efficient

CASE FOLDING

- Applications like IR: reduce all letters to lower case
  - Since users tend to use lower case
  - Possible exception: upper case in mid-sentence?
    - e.g., ***General Motors***
    - ***Fed*** vs. ***fed***
    - ***SAIL*** vs. ***sail***
- For sentiment analysis, MT, Information extraction
  - Case is helpful (***US*** versus ***us*** is important)

LEMMATIZATION

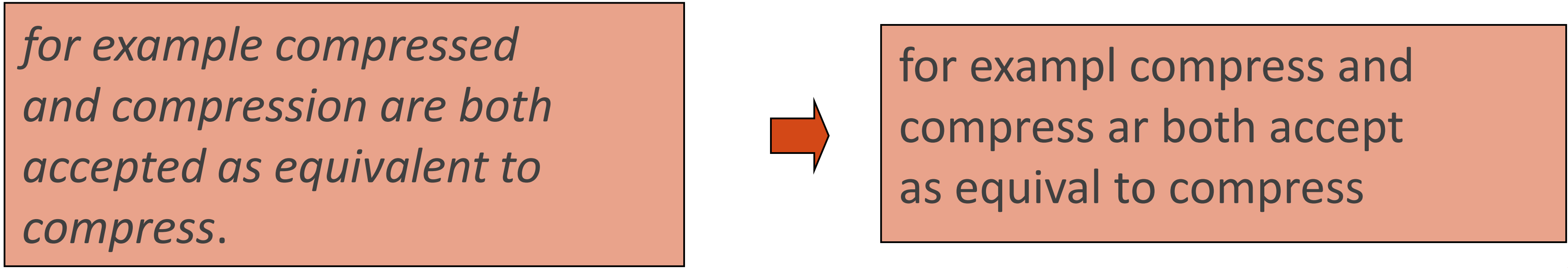
- Reduce inflections or variant forms to base form
  - *am, are, is* → *be*
  - *car, cars, car's, cars'* → *car*
- *the boy's cars are different colors* → *the boy car be different color*
- Lemmatization: have to find correct dictionary headword form
- Machine translation
  - Spanish quiero ('I want'), quieres ('you want') same lemma as querer 'want'

MORPHOLOGY

- **Morphemes:**
  - The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Bits and pieces that adhere to stems
    - Often with grammatical functions

STEMMING

- Reduce terms to their stems in information retrieval
- *Stemming* is crude chopping of affixes
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.



for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

PORTER'S ALGORITHM
THE MOST COMMON ENGLISH STEMMER

Step 1a

sses	→ ss	caresses	→	caress
ies	→ i	ponies	→	poni
ss	→ ss	caress	→	caress
s	→ ø	cats	→	cat

Step 1b

(*v*)ing	→ ø	walking	→	walk
		sing	→	sing
(*v*)ed	→ ø	plastered	→	plaster

…

Step 2 (for long stems)

ational→	ate	relational→		relate
izer→	ize	digitizer	→	digitize
ator→	ate	operator	→	operate

…

Step 3 (for longer stems)

al	→ ø	revival	→	reviv
able	→ ø	adjustable	→	adjust
ate	→ ø	activate	→	activ

…

VIEWING MORPHOLOGY IN A CORPUS
WHY ONLY STRIP –ING IF THERE IS A VOWEL?

(*v*)ing → ø	walking	→	walk
	sing	→	sing

VIEWING MORPHOLOGY IN A CORPUS
WHY ONLY STRIP –ING IF THERE IS A VOWEL?

```
(*v*)ing → ø  walking   → walk
              sing      → sing
```

```
tr -sc 'A-Za-z' '\n' < shakes.txt | grep ’ing$' | sort | uniq -c | sort –nr
```

```
1312 King                    548 being
 548 being                   541 nothing
 541 nothing                 152 something
 388 king                    145 coming
 375 bring                   130 morning
 358 thing                   122 having
 307 ring                    120 living
 152 something               117 loving
 145 coming                  116 Being
 130 morning                 102 going
```

```
tr -sc 'A-Za-z' '\n' < shakes.txt | grep '[aeiou].*ing$' | sort | uniq -c | sort –nr
```

DEALING WITH COMPLEX MORPHOLOGY IS SOMETIMES NECESSARY

- Some languages requires complex morpheme segmentation
  - Turkish
  - Uygarlastiramadiklarimizdanmissinizcasina
  - `(behaving) as if you are among those whom we could not civilize'
  - Uygar `civilized' + las `become'

    + tir `cause' + ama `not able'

    + dik `past' + lar 'plural'

    + imiz 'p1pl' + dan 'abl'

    + mis 'past' + siniz '2pl' + casina 'as if'

BASIC TEXT PROCESSING

Word Normalization and Stemming

LANGUAGE MODELING

Introduction to N-grams

PROBABILISTIC LANGUAGE MODELS

- Today's goal: assign a probability to a sentence
    - Machine Translation:
        - P(**high** winds tonite) > P(**large** winds tonite)
    - Spell Correction
        - The office is about fifteen **minuets** from my house
            - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
    - Speech Recognition
        - P(I saw a van) >> P(eyes awe of an)
    - + Summarization, question-answering, etc., etc.!!

Why?

PROBABILISTIC LANGUAGE MODELING

- Goal: compute the probability of a sentence or sequence of words:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$

- Related task: probability of an upcoming word:

  $P(w_5|w_1,w_2,w_3,w_4)$

- A model that computes either of these:

  $P(W)$ or $P(w_n|w_1,w_2 \ldots w_{n-1})$ is called a **language model**.

- Better: **the grammar** But **language model** or **LM** is standard

HOW TO COMPUTE P(W)

- How to compute this joint probability:
  - P(its, water, is, so, transparent, that)
- Intuition: let’s rely on the Chain Rule of Probability

REMINDER: THE CHAIN RULE

- Recall the definition of conditional probabilities

Rewriting:

- More variables:

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$

- The Chain Rule in General

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

THE CHAIN RULE APPLIED TO COMPUTE JOINT PROBABILITY OF WORDS IN SENTENCE

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P("its water is so transparent") =

P(its) × P(water|its) × P(is|its water)

× P(so|its water is) × P(transparent|its water is so)

HOW TO ESTIMATE THESE PROBABILITIES

- Could we just count and divide?

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- No! Too many possible sentences!
- We'll never see enough data for estimating these

MARKOV ASSUMPTION

Andrei Markov

- Simplifying assumption:

$$P(\text{the} \mid \text{its water is so transparent that}) » P(\text{the} \mid \text{that})$$

- Or maybe

$$P(\text{the} \mid \text{its water is so transparent that}) » P(\text{the} \mid \text{transparent that})$$

MARKOV ASSUMPTION

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

- In other words, we approximate each component in the product

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-k} \ldots w_{i-1})$$

SIMPLEST CASE: UNIGRAM MODEL

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Some automatically generated sentences from a unigram model

fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass

thrift, did, eighty, said, hard, 'm, july, bullish

that, or, limited, the

BIGRAM MODEL

- Condition on the previous word:

$$P(w_i \mid w_1 w_2 \square \;\; w_{i-1}) \text{ » } P(w_i \mid w_{i-1})$$

```
texaco, rose, one, in, this, issue, is, pursuing, growth, in,
a, boiler, house, said, mr., gurria, mexico, 's, motion,
control, proposal, without, permission, from, five, hundred,
fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november
```

N-GRAM MODELS

- We can extend to trigrams, 4-grams, 5-grams
- In general this is an insufficient model of language
  - because language has **long-distance dependencies**:

> "The computer which I had just put into the machine room on the fifth floor crashed."

- But we can often get away with N-gram models

LANGUAGE MODELING

Introduction to N-grams

LANGUAGE MODELING

Estimating N-gram Probabilities

ESTIMATING BIGRAM PROBABILITIES

- The Maximum Likelihood Estimate

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

AN EXAMPLE

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(\texttt{I} \mid \texttt{<s>}) = \tfrac{2}{3} = .67 \qquad P(\texttt{Sam} \mid \texttt{<s>}) = \tfrac{1}{3} = .33 \qquad P(\texttt{am} \mid \texttt{I}) = \tfrac{2}{3} = .67$$

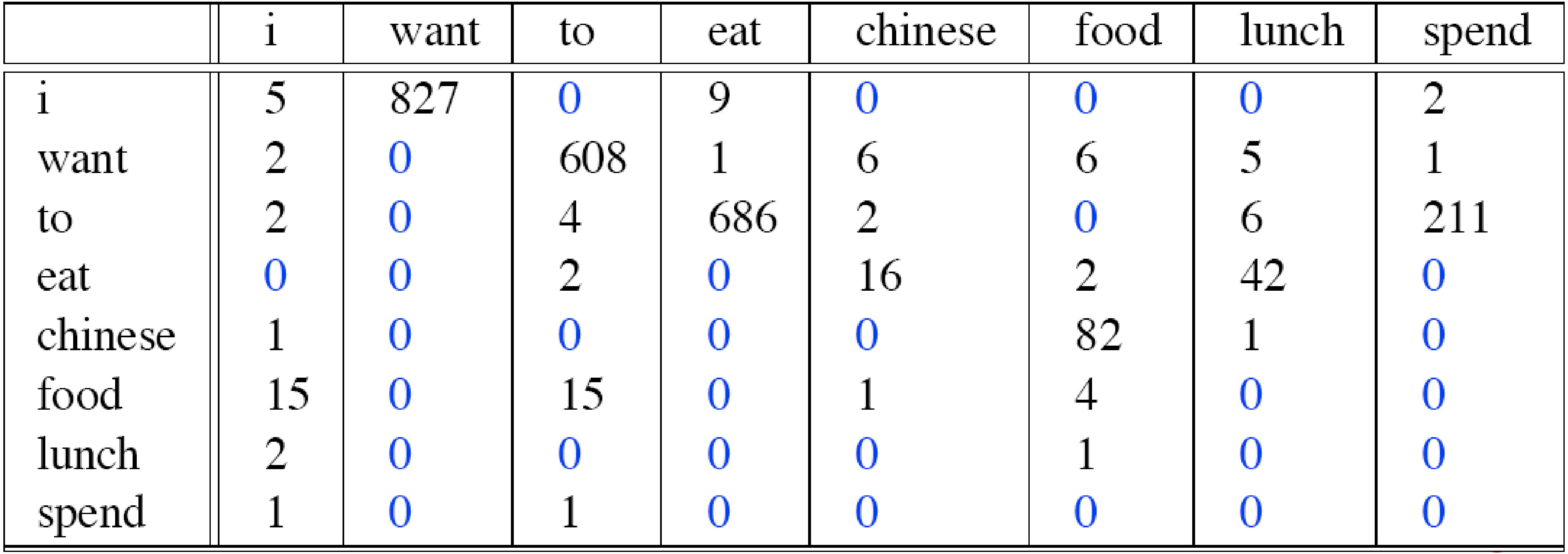
$$P(\texttt{</s>} \mid \texttt{Sam}) = \tfrac{1}{2} = 0.5 \qquad P(\texttt{Sam} \mid \texttt{am}) = \tfrac{1}{2} = .5 \qquad P(\texttt{do} \mid \texttt{I}) = \tfrac{1}{3} = .33$$

MORE EXAMPLES:
BERKELEY RESTAURANT PROJECT SENTENCES

- can you tell me about any good cantonese restaurants close by
- mid priced thai food is what i'm looking for
- tell me about chez panisse
- can you give me a listing of the kinds of food that are available
- i'm looking for a good place to eat breakfast
- when is caffe venezia open during the day

RAW BIGRAM COUNTS

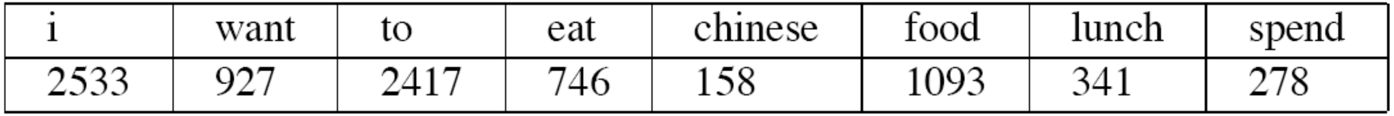
- Out of 9222 sentences



	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

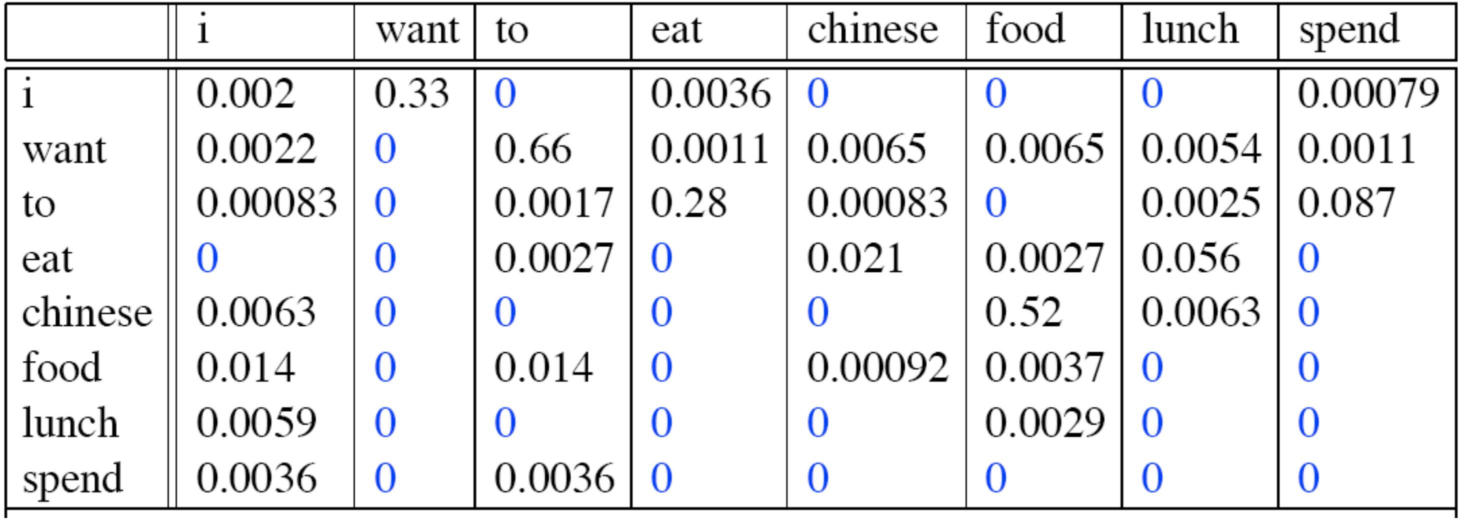
RAW BIGRAM PROBABILITIES

- Normalize by unigrams:



i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Result



	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

BIGRAM ESTIMATES OF SENTENCE PROBABILITIES

P(<s> I want english food </s>) =

P(I|<s>)

× P(want|I)

× P(english|want)

× P(food|english)

× P(</s>|food)

= .000031

WHAT KINDS OF KNOWLEDGE?

- P(english|want) = .0011
- P(chinese|want) = .0065
- P(to|want) = .66
- P(eat | to) = .28
- P(food | to) = 0
- P(want | spend) = 0
- P (i | <s>) = .25

PRACTICAL ISSUES

- We do everything in log space
  - Avoid underflow
  - (also adding is faster than multiplying)

$$\log(p_1 \text{´}\ p_2 \text{´}\ p_3 \text{´}\ p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

LANGUAGE MODELING

Estimating N-gram Probabilities