CS325 Artificial Intelligence
Natural Language Processing I (Ch. 22)

Dr. Cengiz Günay, Emory Univ.

Spring 2013

AI in Natural Language Processing (NLP)

What's NLP?

AI in Natural Language Processing (NLP)

What's NLP?

- Computers understanding our languages: English, French, Japanese, ...

AI in Natural Language Processing (NLP)

What's NLP?

- Computers understanding our languages: English, French, Japanese, ...

Why?

AI in Natural Language Processing (NLP)

What's NLP?

- Computers understanding our languages: English, French, Japanese, ...

Why?

- We can talk to the computer

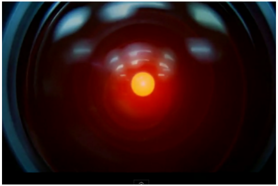
AI in Natural Language Processing (NLP)

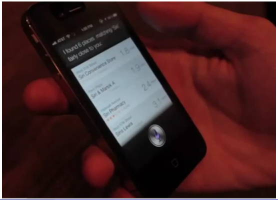
What's NLP?

- Computers understanding our languages: English, French, Japanese, ...

Why?

- We can talk to the computer
- It can talk to us, too





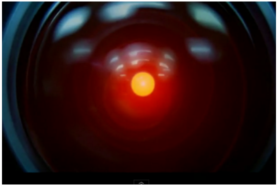
AI in Natural Language Processing (NLP)

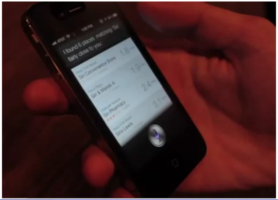
What's NLP?

- Computers understanding our languages: English, French, Japanese, ...

Why?

- We can talk to the computer
- It can talk to us, too
- And it can read our stuff





Entry/Exit Surveys

Exit survey: Robotics II – Navigation

- Why do we normalize particle weights? Where are they used next?
- How can we force a robot whether or not to choose actions like taking a left turn?

Entry survey: Natural Language Processing I (0.25 pts)

- Give some examples, other than what I showed, where NLP would be useful.
- Explain briefly how the spam filter in the machine learning lecture worked.

What Can We Do With NLP?

What NLP task we can do with the following?

Classification:

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering:

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

What Can We Do With NLP?

What NLP task we can do with the following?

Classification:	Spam vs. Ham
Clustering:	News articles, emails, . . .
Spelling:	

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: Atuo-crorect

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: ~~Atuo-crorect~~,

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: ~~Atuo-crorect~~, auto-correct

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking:

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, ...

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking: Read user reviews.

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, ...

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking: Read user reviews.

Information retrieval:

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking: Read user reviews.

Information retrieval: Search engines.

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking: Read user reviews.

Information retrieval: Search engines.

Answering questions:

What Can We Do With NLP?

What NLP task we can do with the following?

Classification:	Spam vs. Ham
Clustering:	News articles, emails, . . .
Spelling:	~~Atuo-crorect~~, auto-correct
Product ranking:	Read user reviews.
Information retrieval:	Search engines.
Answering questions:	IBM's Watson.

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, . . .

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking: Read user reviews.

Information retrieval: Search engines.

Answering questions: IBM's Watson.

Translation:

What Can We Do With NLP?

What NLP task we can do with the following?

Classification:	Spam vs. Ham
Clustering:	News articles, emails, ...
Spelling:	~~Atuo-crorect~~, auto-correct
Product ranking:	Read user reviews.
Information retrieval:	Search engines.
Answering questions:	IBM's Watson.
Translation:	Google translate, Altavista Babelfish.

What Can We Do With NLP?

What NLP task we can do with the following?

Classification:	Spam vs. Ham
Clustering:	News articles, emails, ...
Spelling:	~~Atuo-crorect~~, auto-correct
Product ranking:	Read user reviews.
Information retrieval:	Search engines.
Answering questions:	IBM's Watson.
Translation:	Google translate, Altavista Babelfish.
Speech recognition:	

What Can We Do With NLP?

What NLP task we can do with the following?

Classification:	Spam vs. Ham
Clustering:	News articles, emails, . . .
Spelling:	~~Atuo-crorect~~, auto-correct
Product ranking:	Read user reviews.
Information retrieval:	Search engines.
Answering questions:	IBM's Watson.
Translation:	Google translate, Altavista Babelfish.
Speech recognition:	Diction programs, Siri.
Learning:	

What Can We Do With NLP?

What NLP task we can do with the following?

Classification: Spam vs. Ham

Clustering: News articles, emails, ...

Spelling: ~~Atuo-crorect~~, auto-correct

Product ranking: Read user reviews.

Information retrieval: Search engines.

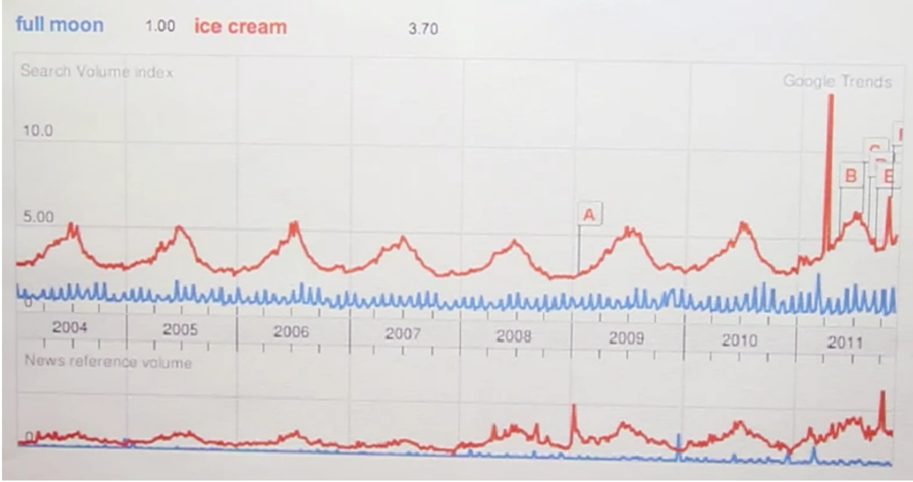
Answering questions: IBM's Watson.

Translation: Google translate, Altavista Babelfish.

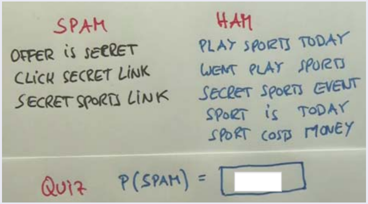
Speech recognition: Diction programs, Siri.

Learning: Tap into the world's knowledge...

Learning From Language



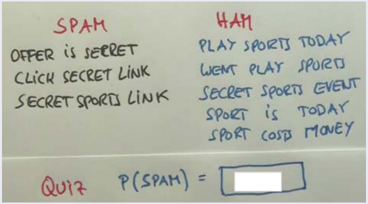
Two methods:



	Model	How?	Construct?
	Probabilistic		

How Can We Understand Language?

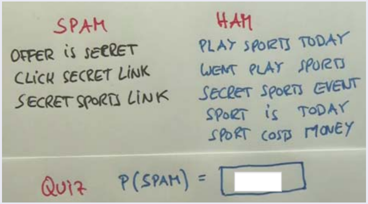
Two methods:



Model	How?	Construct?
Probabilistic	Word-based	

How Can We Understand Language?

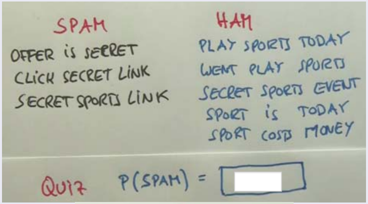
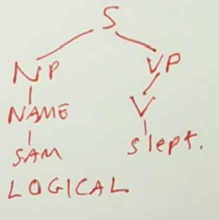
Two methods:



Model	How?	Construct?
Probabilistic	Word-based	Learned

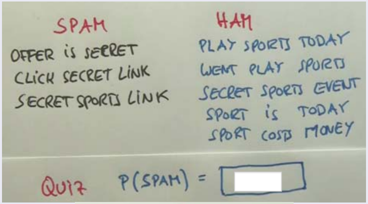
How Can We Understand Language?

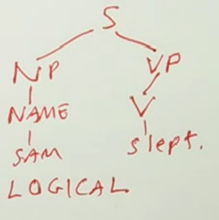
Two methods:

	Model	How?	Construct?
	Probabilistic	Word-based	Learned
	Logical		

How Can We Understand Language?

Two methods:

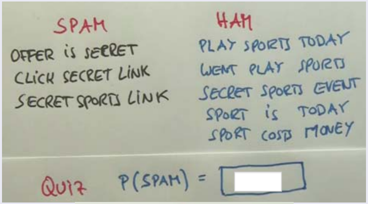


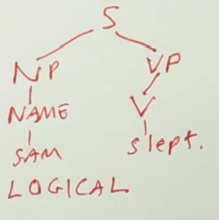


Model	How?	Construct?
Probabilistic	Word-based	Learned
Logical	Grammar	

How Can We Understand Language?

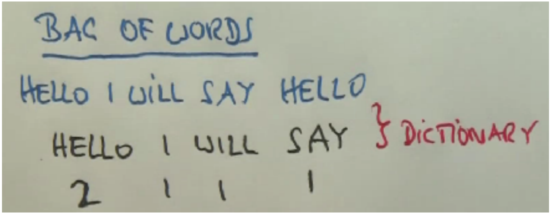
Two methods:





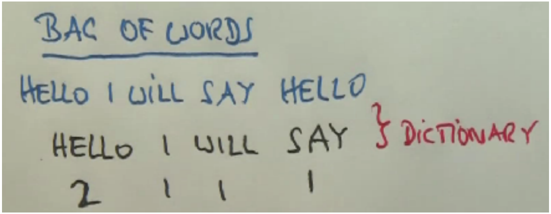
Model	How?	Construct?
Probabilistic	Word-based	Learned
Logical	Grammar	Programmed

Remember Bag of Words?



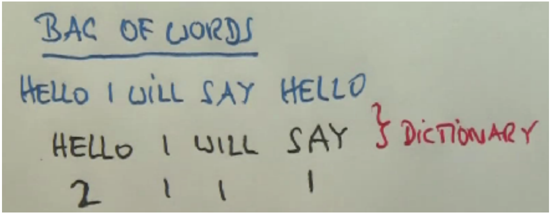
$P(\text{Hello}) = ?$

Remember Bag of Words?



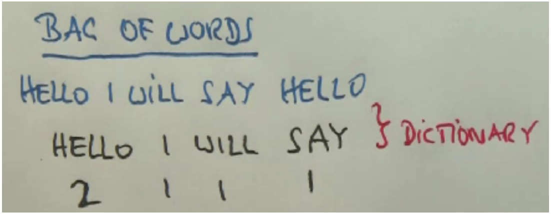
$P(\text{Hello}) = \frac{2}{5}$

Remember Bag of Words?



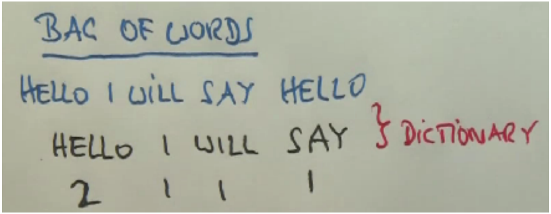
$P(\text{Hello}) = \frac{2}{5}$
$P(\text{I}) = ?$

Remember Bag of Words?



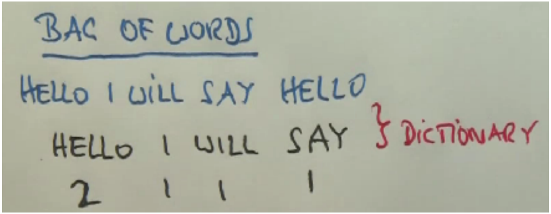
$P(\text{Hello}) = \frac{2}{5}$
$P(\text{I}) = \frac{1}{5}$

Remember Bag of Words?



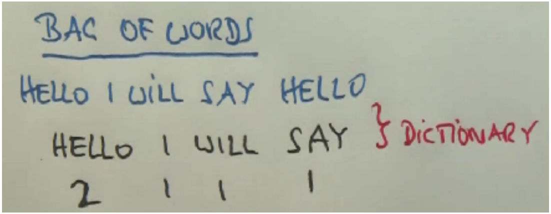
$P(\text{Hello}) = \frac{2}{5}$
$P(\text{I}) = \frac{1}{5}$
$= P(\text{Will}) = P(\text{Say})$

Remember Bag of Words?



Words are independent?

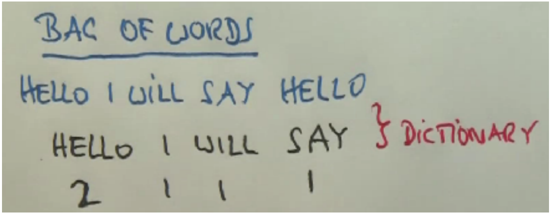
$$P(\text{Hello}) = \frac{2}{5}$$
$$P(\text{I}) = \frac{1}{5}$$
$$= P(\text{Will}) = P(\text{Say})$$



$$P(\text{Hello}) = \frac{2}{5}$$
$$P(\text{I}) = \frac{1}{5}$$
$$= P(\text{Will}) = P(\text{Say})$$

Words are independent? Called **unigram** or **1-gram**:

Remember Bag of Words?



$$P(\text{Hello}) = \frac{2}{5}$$
$$P(\text{I}) = \frac{1}{5}$$
$$= P(\text{Will}) = P(\text{Say})$$

Words are independent? Called **unigram** or **1-gram**:

$$P(w_1, w_2, \ldots, w_n) \quad = \quad \prod_i P(w_i)$$

Can we get more from Bayes?

Distinguish between:

"I will say hello"

"I hello say will"

Can we get more from Bayes?

Distinguish between:

"I will say hello"

"I hello say will"

$$P(\text{"hello"} \,|\, \text{"I will say"}) \overset{>}{<} P(\text{"will"} \,|\, \text{"I hello say"})$$

Can we get more from Bayes?

Distinguish between:

"I will say hello"

"I hello say will"

$$P(\text{"hello"} \mid \text{"I will say"}) > P(\text{"will"} \mid \text{"I hello say"})$$

Can we get more from Bayes?

Distinguish between:

"I will say hello"

"I hello say will"

$$P(\text{"hello"} \,|\, \text{"I will say"}) > P(\text{"will"} \,|\, \text{"I hello say"})$$

Words dependent on previous words: called ***N*-gram**

Can we get more from Bayes?

Distinguish between:

"I will say hello"

"I hello say will"

$$P(\text{"hello"}\,|\,\text{"I will say"}) > P(\text{"will"}\,|\,\text{"I hello say"})$$

Words dependent on previous words: called ***N*-gram**

$$\begin{aligned} P(w_1, w_2, \ldots, w_n) &= P(w_{1:n}) \\ &= \prod_i P(w_i | w_{1:(i-1)}) \end{aligned}$$

Must Remember All Words That Came Before?

Thomas Bayes was the son of London Presbyterian minister Joshua Bayes[4] and was possibly born in Hertfordshire.[5] He came from a prominent non conformist family from Sheffield. In 1719, he enrolled at the University of Edinburgh to study logic and theology. On his return around 1722, he assisted his father at the latter's non-conformist chapel in London before moving to Tunbridge Wells, Kent around 1734. There he became minister of the Mount Sion chapel, until 1752.[6]

$$P(\text{"1752"} \mid \text{"Thomas Bayes} \ldots\text{"}) = ?$$

Must Remember All Words That Came Before?

Thomas Bayes was the son of London Presbyterian minister Joshua Bayes[4] and was possibly born in Hertfordshire.[5] He came from a prominent non conformist family from Sheffield. In 1719, he enrolled at the University of Edinburgh to study logic and theology. On his return around 1722, he assisted his father at the latter's non-conformist chapel in London before moving to Tunbridge Wells, Kent around 1734. There he became minister of the Mount Sion chapel, until 1752.[6]

$$P(\text{"}1752\text{"} \,|\, \text{"Thomas Bayes} \ldots \text{"}) = ?$$

Markov assumption: Only remember last N words: N-gram.

Must Remember All Words That Came Before?

Thomas Bayes was the son of London Presbyterian minister Joshua Bayes[4] and was possibly born in Hertfordshire.[5] He came from a prominent non conformist family from Sheffield. In 1719, he enrolled at the University of Edinburgh to study logic and theology. On his return around 1722, he assisted his father at the latter's non-conformist chapel in London before moving to Tunbridge Wells, Kent around 1734. There he became minister of the Mount Sion chapel, until 1752.[6]

$$P(\text{"1752"} \,|\, \text{"Thomas Bayes \ldots"}) = ?$$

Markov assumption: Only remember last N words: N-gram.

$$P(w_{1:k}) = \prod_{i}^{k} P(w_i | w_{(i-N):(i-1)})$$

Let's Read Shakespeare... In Unigram

Unigram=1-gram

n=1: : more by and that .

n=1: volumnius ears stealing very am , remember go quality in error , my this wherefore jessica talk'd me an first prove maid's all .

n=1: : while leaping-houses ear i !

n=1: thou hurt , we ; ?

n=1: a us if at the undiscovered thou o'erthrown he'll this theft issu'd !

n=1: shut cur an court to again rock call'd triumvirate best she and before will .

n=1: the slept , they the , the conjures for me eyes !

n=1: on ; no of aweary sea-farer those for as yield the creatures be not a , but the did comes 'tis ; can have , allowance .

n=1: , but i speak my dear .

n=1: we home one , see of : , will should brave , as , or kind fasten'd steal near man's i shall , if their our , stay , know age ; it , is , and likewise .

Shakespeare In Bigram

$N = 2$: bigram

n=2: if thou sober-suited matron , prick me be call'd a bastard .

n=2: peasant swain !

n=2: 'tis but a duck again give you do beseech you , and the king and palmy state to you sit sore eye of your name ?

n=2: marry them : 'tis but , biondello , and liberal opposition .

n=2: come the city here to make her .

n=2: i have lived in each , salanio ?

n=2: hear you come to think ?

n=2: hark ye , then : but to reprobation .

n=2: what should be ta'en a fool ?

n=2: did these lovers into your city call us lord , rather than want a colour that i had in it , sir , by sea and land , as it would not change this purpose cool : i will look further into't ; and every one an empty coffer : lay thine ear to hear from me the way of argument .

Shakespeare In Trigram

$N = 3$: trigram

n=3: the gods to send the companion a better husband .

n=3: come , grey of northumberland .

n=3: how do you take pains to con them by the inward motion to deliver us !

n=3: let your lady being so easy and so we'll leave a thousand-fold more bitter than 'tis sweet at first .

n=3: i will never yield .

n=3: i cannot be mine , are well .

n=3: little pretty ones !

n=3: you are always my good friends .

n=3: i would learn of noble edward's sons , what thing , avoid !

n=3: give me audience : if once i encounter'd him , being the mered question .

Shakespeare In 4-gram

n=4: my master knows not of your wrong .

n=4: the augurer tells me we shall have great store of wedding cheer ; but so it is , i came with no ill intent , for to that the working of your own cause .

n=4: this tarsus , o'er which i have told my neighbour how you have dealt for him ; or ere i journey to your father's choice , you can produce acquittances for such a business give me leave .

n=4: i should sin to think , that had put such difference betwixt their two estates ; love no god , that in your countenance which i would fain see it once , and that my path were even to the frozen ridges of the alps , or any taint of vice whose strong corruption inhabits our frail blood .

n=4: will you shog off ?

n=4: i know my duty .

n=4: who am i , and i his fate .

n=4: here's my glove : give me some little breath , some pause , dear lord , before i speak that you make known it is no matter , sir : the rascal's drunk .

Shakespeare *N*-gram Quiz

1 2 3	my duty, and tell us what occasion now, what's become of me.
1 2 3	clap begetting home prove and you unless he will,, your passages,..
1 2 3	i was affianced to england; if my will live to you can do no countermand, shall have done 't.
1 2 3	exit, pursued by a bear.
1 2 3	under this inconvenience, the which drives; for, and kneel thou melancholy.
1 2 3	, but her hours report have one, ask may some.
1 2 3	woe is me to remember that the bastard; take't up, i was not in the name of king henry.
1 2 3	in verona, not very beastly, come away!
1 2 3	he cannot be measur'd rightly, your inclining cannot be a mock: i say he shall be mine.
1 2 3	lose, wife devil the.

Find:

- 1 real quote
- 3x unigram picks
- 3x bigram picks
- 3x trigram picks

Shakespeare *N*-gram Quiz

1 2 3	my duty. and tell us what occasion now. what's become of me.
1 2 3	clap begetting home prove and you unless he will,. your passages,.
1 2 3	i was affianced to england. if my will live to you can do no countermand. shall have done 't.
1 2 3	exit, pursued by a bear.
1 2 3	under this inconvenience, the which drives; for, and kneel thou melancholy.
1 2 3	, but her hours report have one, ask may some.
1 2 3	woe is me to remember that the bastard; take't up, i was not in the name of king henry.
1 2 3	in verona, not very beastly, come away!
1 2 3	he cannot be measur'd rightly. your inclining cannot be a mock; i say he shall be mine.
1 2 3	lose, wife devil the.

Find:

- 1 real quote
- 3x unigram picks
- 3x bigram picks
- 3x trigram picks

Bigram Probability Question

$P(\hat{}\,\mathrm{woe\ is\ me}|\hat{}) = ?$

Given that:

$\hat{}$: symbol showing start of sentence

$P(\mathrm{woe}_i|\hat{}_{i-1}) = .0002$

$P(\mathrm{is}_i|\mathrm{woe}_{i-1}) = .07$

$P(\mathrm{me}_i|\mathrm{is}_{i-1}) = .0005$

Bigram Probability Question

$P(\hat{}\,\mathrm{woe\ is\ me}|\hat{}\,) = .0002 \times .07 \times .0005 = \mathbf{7 \times 10^{-9}}$

Given that:

$\hat{}$: symbol showing start of sentence

$P(\mathrm{woe}_i|\hat{}_{i-1}) = .0002$

$P(\mathrm{is}_i|\mathrm{woe}_{i-1}) = .07$

$P(\mathrm{me}_i|\mathrm{is}_{i-1}) = .0005$

Other Tricks

Stationarity assumption: Context doesn't change over time.

Other Tricks

Stationarity assumption: Context doesn't change over time.

Smoothing: Remember Laplace smooting?

Other Tricks

Stationarity assumption: Context doesn't change over time.

Smoothing: Remember Laplace smooting?

Hidden variables: E.g., identify what a "noun" is.

Other Tricks

Stationarity assumption: Context doesn't change over time.

Smoothing: Remember Laplace smooting?

Hidden variables: E.g., identify what a "noun" is.

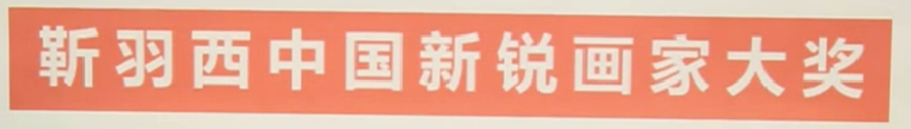
Use abstractions: Group "New York City", or just look at letters.

Smaller Than Words?

What if we cannot distinguish words?

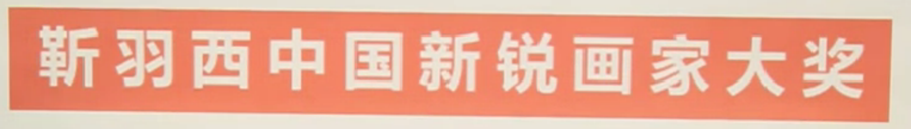
Smaller Than Words?

What if we cannot distinguish words?



Smaller Than Words?

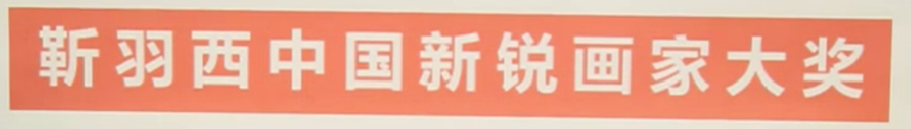
What if we cannot distinguish words?



English: “choosespain.com”
“Choose Spain” OR
“Chooses Pain”?

Smaller Than Words?

What if we cannot distinguish words?

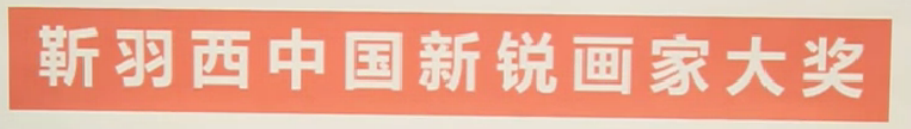


English: “choosespain.com”
“Choose Spain” OR
“Chooses Pain”?

Segmentation: Dividing into words.

Smaller Than Words?

What if we cannot distinguish words?



English: “choosespain.com”
“Choose Spain” OR
“Chooses Pain”?

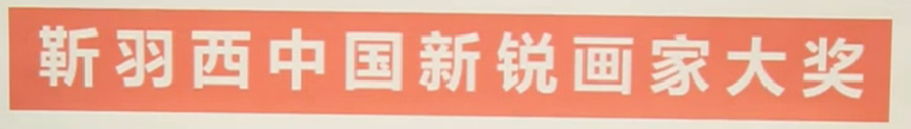
Segmentation: Dividing into words.

Use Bayes again:

$$s^* = \max P(w_{1:n}) = \max \prod_i P(w_i | w_{1:i})$$

Smaller Than Words?

What if we cannot distinguish words?



English: "choosespain.com"
"Choose Spain" OR
"Chooses Pain"?

Segmentation: Dividing into words.

Use Bayes again:

$$s^* = \max P(w_{1:n}) = \max \prod_i P(w_i | w_{1:i})$$

Or Markov assumption (e.g., unigram):

$$s^* = \max \prod_i P(w_i)$$

Segmentation Complexity

$$s^* = \max \prod_i P(w_i)$$

What's the complexity of segmenting: **"nowisthetime"**?

Segmentation Complexity

$$s^* = \max \prod_i P(w_i)$$

What's the complexity of segmenting: **"nowisthetime"**?

1. $n - 1$
2. $(n - 1)^2$
3. $(n - 1)!$
4. 2^{n-1}
5. $(n - 1)^n$

Segmentation Complexity

$$s^* = \max \prod_i P(w_i)$$

What's the complexity of segmenting: **"nowisthetime"**?

1. $n - 1$
2. $(n - 1)^2$
3. $(n - 1)!$
4. $\mathbf{2^{n-1}}$
5. $(n - 1)^n$

Segmentation Complexity

$$s^* = \max \prod_i P(w_i)$$

What's the complexity of segmenting: **"nowisthetime"**?

1. $n - 1$
2. $(n-1)^2$
3. $(n-1)!$
4. $\mathbf{2^{n-1}}$
5. $(n-1)^n$

Solution: Separate each character with $n - 1$ divisions, form words by whether division exists or not.

Reducing Segmentation Complexity

Exploit independence: **"nowisthetime"**?

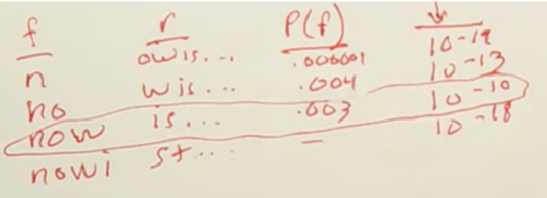
Reducing Segmentation Complexity

Exploit independence: **"nowisthetime"**?
Divide into first, f, and recurse for rest, r:

$$s^* = \max_{s=f+r} P(f) \cdot s^*(r)$$

Exploit independence: **"nowisthetime"**?
Divide into first, f, and recurse for rest, r:

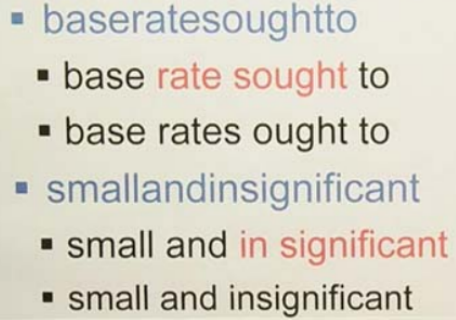
$$s^* = \max_{s=f+r} P(f) \cdot s^*(r)$$



Gives **99% accuracy** and **easy implementation**!

Segmentation Problems

- baseratesoughtto
  - base rate sought to
  - base rates ought to
- smallandinsignificant
  - small and in significant
  - small and insignificant



How can we improve?

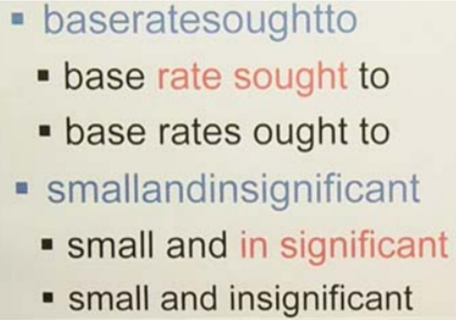
1. More Data
2. Markov
3. Smoothing

- baseratesoughtto
  - base rate sought to
  - base rates ought to
- smallandinsignificant
  - small and in significant
  - small and insignificant

How can we improve?

1. More Data
2. **Markov**
3. Smoothing

Segmentation Problems



Need to get the **context**.

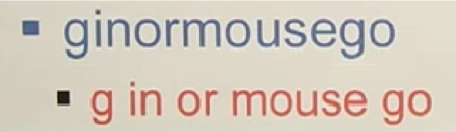
How can we improve?

1. More Data
2. Markov
3. Smoothing

Segmentation Problems (2)

- ginormousego
  - g in or mouse go

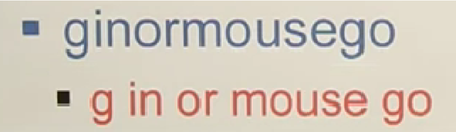
Segmentation Problems (2)



How can we improve?

1. More Data
2. Markov
3. Smoothing

Segmentation Problems (2)



How can we improve?

1. More Data
2. Markov
3. Smoothing

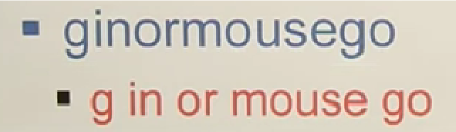
Segmentation Problems (2)

- ginormousego
  - g in or mouse go

How can we improve?

1. **More Data**
2. Markov
3. **Smoothing**

Segmentation Problems (2)

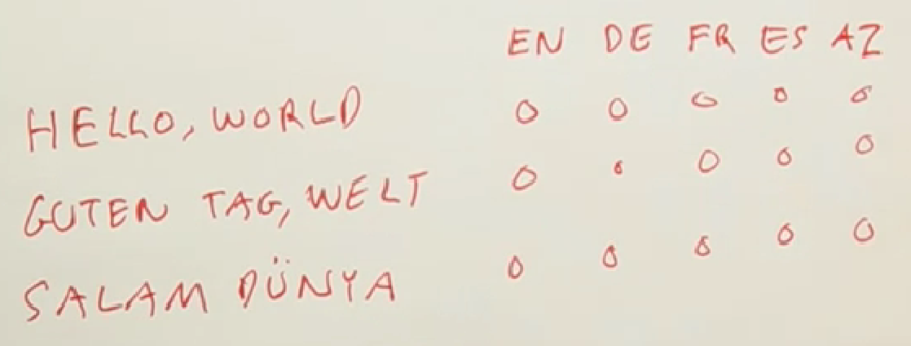


How can we improve?

1. More Data
2. Markov
3. Smoothing

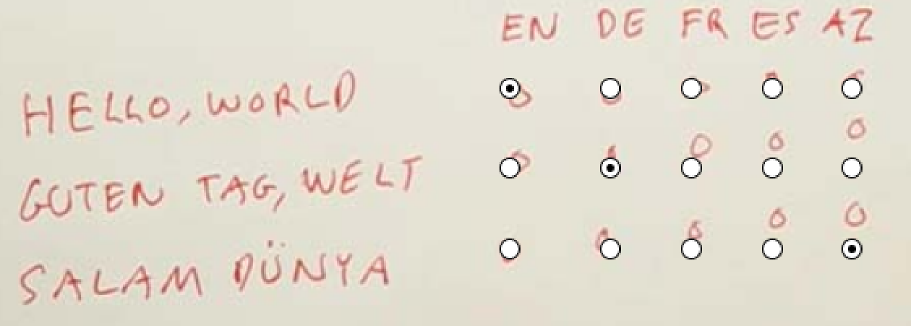
Need to **know more words**.

Language identification?



What Else Can We Do with Letters?

Language identification?



Bigram Recognition with Letters

#	A	B	C
1	TH	EN	IN
2	TE	ER	AN
3	OU	CH	ƏR
4	AN	DE	LA
5	ER	EI	IR
6	IN	IN	AR
English?	o	o	o
German?	o	o	o
Azerbijani?	o	o	o

Bigram Recognition with Letters

#	A	B	C
1	TH	EN	IN
2	TE	ER	AN
3	OU	CH	ƏR
4	AN	DE	LA
5	ER	EI	IR
6	IN	IN	AR
English?	◉	○	○
German?	○	◉	○
Azerbijani?	○	○	◉

Trigram Recognition with Letters

#	A	B	C
P(the)	1.1%	.03%	.00%
P(der)	.06%	.68%	.00%
P(rba)	.00%	.01%	.53%
English?	o	o	o
German?	o	o	o
Azerbijani?	o	o	o

Trigram Recognition with Letters

#	A	B	C
P(the)	1.1%	.03%	.00%
P(der)	.06%	.68%	.00%
P(rba)	.00%	.01%	.53%
English?	◉	○	○
German?	○	◉	○
Azerbijani?	○	○	◉

Trigram Recognition with Letters

#	A	B	C
P(the)	1.1%	.03%	.00%
P(der)	.06%	.68%	.00%
P(rba)	.00%	.01%	.53%
English?	◉	○	○
German?	○	◉	○
Azerbijani?	○	○	◉

#	A	B	C
P(the)	1.1%	.03%	.00%
P(der)	.06%	.68%	.00%
P(rba)	.00%	.01%	.53%
English?	◉	○	○
German?	○	◉	○
Azerbijani?	○	○	◉

99% accuracy from trigrams!

Can We Identify Categories Too?

People	Places	Drugs
Steve Jobs	San Francisco	Lipitor
Bill Gates	Palo Alto	Prevacid
Andy Grove	Stern Grove	Zoloft
Larry Page	San Mateo	Zocor
Andrew Ng	Santa Cruz	Plavix
Jennifer Widom	New York	Protonix
Daphne Koller	New Jersey	Celebrex
Noah Goodman	Jersey City	Zyrtec
Julie Zelinski	South San Francisco	Aggrenox

Can We Identify Categories Too?

People	Places	Drugs
Steve Jobs	San Francisco	Lipitor
Bill Gates	Palo Alto	Prevacid
Andy Grove	Stern Grove	Zoloft
Larry Page	San Mateo	Zocor
Andrew Ng	Santa Cruz	Plavix
Jennifer Widom	New York	Protonix
Daphne Koller	New Jersey	Celebrex
Noah Goodman	Jersey City	Zyrtec
Julie Zelinski	South San Francisco	Aggrenox

Text classification

Text Classification

What algorithms can we use?

Naive Bayes:

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor:

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

Support Vector Machines:

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

Support Vector Machines: Supervised learning

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

Support Vector Machines: Supervised learning

Regression:

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

Support Vector Machines: Supervised learning

Regression: Prediction

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

Support Vector Machines: Supervised learning

Regression: Prediction

Zip:

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

k-Nearest Neighbor: Similar words

Support Vector Machines: Supervised learning

Regression: Prediction

Zip: What??

Text Classification

What algorithms can we use?

Naive Bayes: Spam vs. Ham

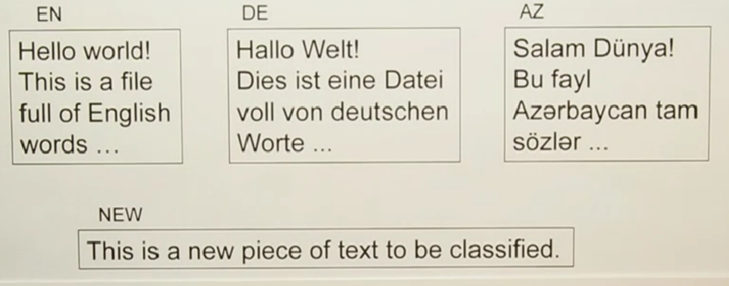
k-Nearest Neighbor: Similar words

Support Vector Machines: Supervised learning

Regression: Prediction

Zip: What??

Compression As Classifier



Compression As Classifier

EN

Hello world!
This is a file
full of English
words ...

DE

Hallo Welt!
Dies ist eine Datei
voll von deutschen
Worte ...

AZ

Salam Dünya!
Bu fayl
Azərbaycan tam
sözlər ...

NEW

This is a new piece of text to be classified.

```
(echo `cat new EN | gzip | wc –c` EN; \
 echo `cat new DE | gzip | wc –c` DE; \
 echo `cat new AZ | gzip | wc –c` AZ) \
| sort –n | head -1
```

Spelling Correction

Correction, c, for word, w:

$$c^* = \max_c P(c|w)$$

Spelling Correction

Correction, c, for word, w:

$$c^* = \max_c P(c|w)$$

Use Bayes Rule:

$$c^* = \max_c P(w|c)P(c)$$

where

$P(c)$ from data counts

$P(w|c)$ from spelling correction data

Spelling Correction Data

C; W, w

$P(w|c)$

$P(pluse|pulse)$

pulse: pluse
elegant: elagent, elligit
second: secand, sexeon, secund, seconnd, seond, sekon
sailed: saled, saild
blouse: boludes
thunder: thounder
cooking: coking, chocking, kooking, cocking
fossil: fosscil

Spelling Correction Example

$w = \text{"thew"}$ $P(w|c) \cdot P(c)$

w	c	$w \mid c$	$P(w \mid c)$	$P(c)$	$10^9 P(w \mid c) P(c)$
thew	the	ew | e	.000007	.02	144.
thew	thew		.95	.00000009	90.
thew	thaw	e | a	.001	.0000007	0.7
thew	threw	h | hr	.000008	.000004	0.03
thew	thwe	ew | we	.000003	.00000004	0.0001