Language Models and Interfaces 2015

Dr Ivan Titov

ILLC, Universiteit van Amsterdam

Administrative issues

Meetings:	2 x 2 hours lectures + 2 x 2 Practical hours
Examination:	Programming assignments, reports and exams
Web page	Reachable via Blackboard, lectures online after the class

Main Text Books and References

Book: D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing*. Computational Linguistics, and Speech Recognition. 2008 (Second Edition).

Smoothing: J. Goodman and S. Chen. *An empirical study of smoothing techniques for language modeling*. Tech. report TR-10-98, Harvard University, August 1998. `http://research.microsoft.com/~joshuago/`

Extra: Ch. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA: May 1999.

Structure of the 1^{st} lecture

- What this course is about?
- Main NLP issues: Ambiguity and Robustness
- Why *probabilistic* models for NLP?
- What are probabilistic models (general)?

- Basic Probability Theory Reminder.

Artificial Intelligence: Goals with Natural Language

- **Extract info:** Systems that extract information from textual or spoken media.
  Examples: inf retrieval, inf extraction, data mining etc.
- **Text2Text:** Systems that transform text/speech to text/speech
  Examples: translation systems, summarization, dictation, reading etc.
- **Dialog:** Systems that communicate with people through language.
  Examples: dialog systems

Question: how does language play a role in these systems?

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Named entity recognition, co-reference (pronoun) resolution

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

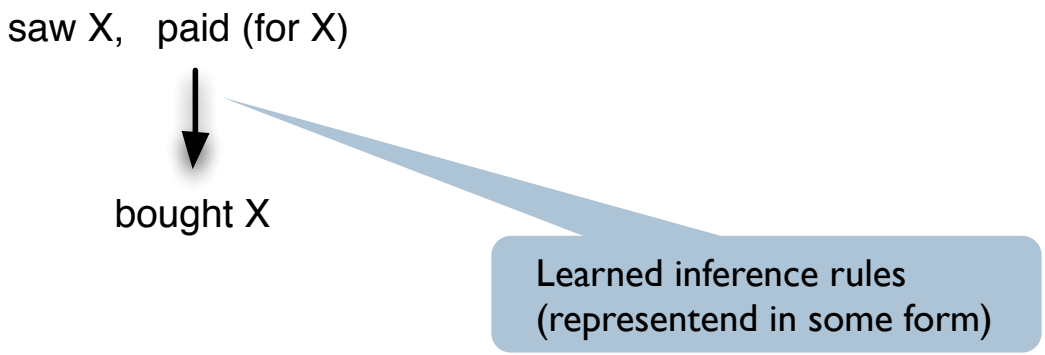
1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Syntactic parsing, shallow semantic analysis (argument identification)

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.



1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example task

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

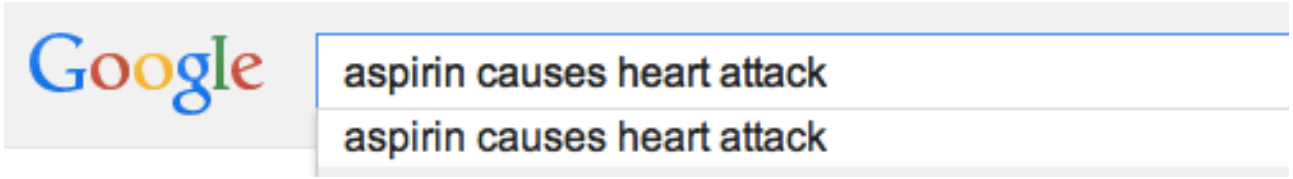
1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Requires inference / Reasoning

But it does not necessary imply the need for formal logical inference

This type of natural language understanding is generally beyond what current systems are capable

Inference in information retrieval



Page 1

… Non-steroidal anti-inflammatory drugs have been shown to increase risk of Myocardial infarction…

Page 2

Nonsteroidal anti-inflammatory drugs (NSAIDs) are a class of drugs which include ibuprofen, **aspirin**, naproxin ...

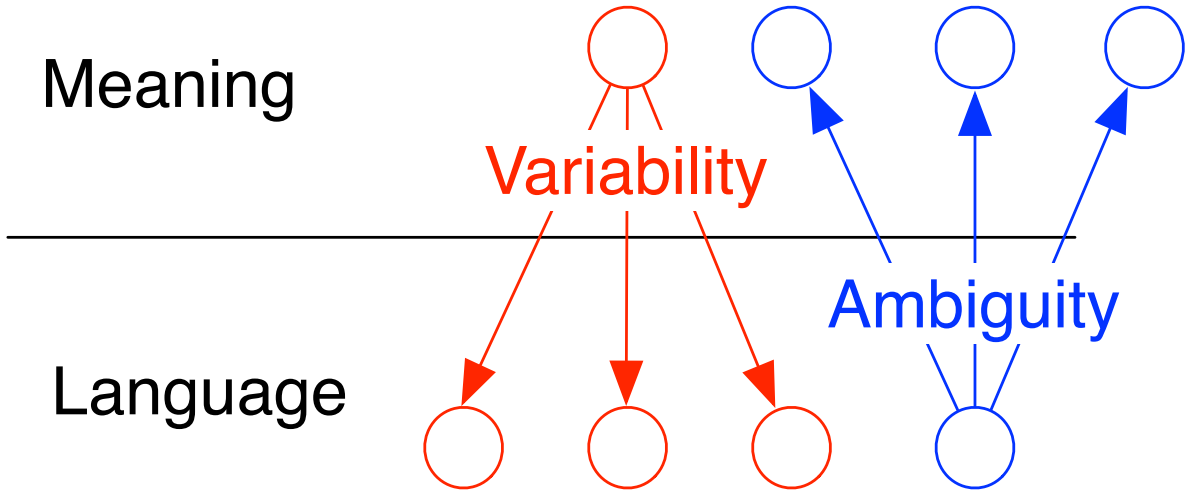
Page 3

Myocardial infractions, also known as **heart attack**, ...

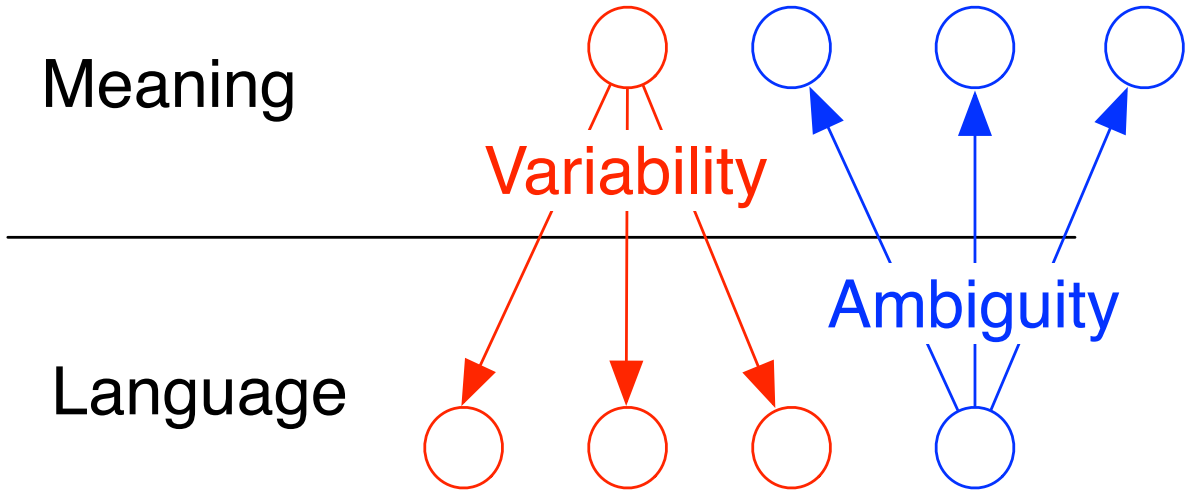
The recent Knowledge Graph initiative looks into this direction

Cross-documents inference is generally beyond reach of IR systems

Why understanding language is hard?



Why understanding language is hard?



Variability:

He drew the house

He made a sketch of the house

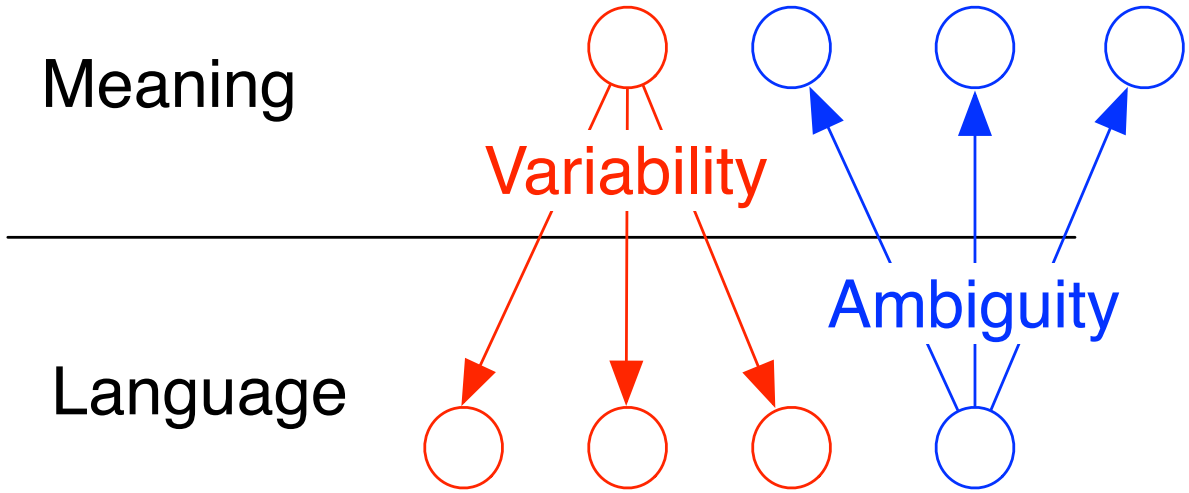
He showed me his drawing of the house

He portrayed the house in his paintings

He drafted the house in his sketchbook

...

Why understanding language is hard?



Ambiguity:

She drew a picture of herself ~ *sketched, made a drawing of*

A cart drawn by two horses… ~ *pulled*

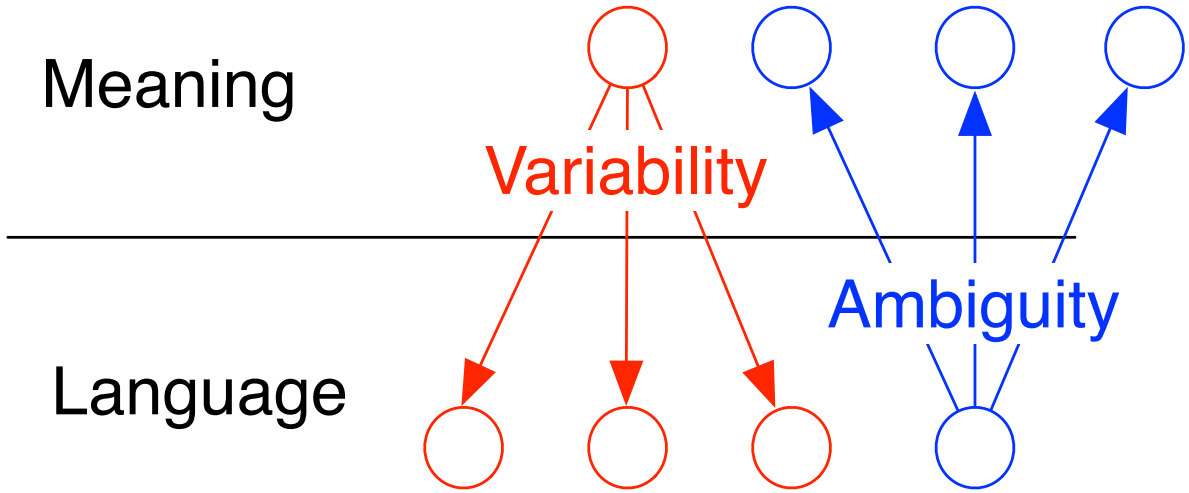
He drew crowds wherever he went … ~ *attracted*

The driver slowed as he drew even with me ~ *proceeded*

The officer drew a gun and pointed it at … ~ *took out, produced*

..

Why understanding language is hard?



Ambiguity:

She drew a picture of herself ~ *sketched, made a drawing of*

A cart drawn by two horses… ~ *pulled*

He drew crowds wherever he went … ~ *attracted*

The driver slowed as

The officer drew a gu

..

In this class we will mostly deal with ambiguity rather than variability

Example system

A computer provides information about train schedules:

C: Good evening. How can I help you?

U: I want to travel to Utrecht. Eh... from Amsterdam tomorrow evening.

C: What time do you want to arrive in Utrecht?

U: I want to depart at around half eight.

C: There is a train at seven thirty six from Amsterdam CS, arriving at seven fifty six in Utrecht CS. Is that suitable for you?

⋮

What components we need for this system?
What problems can we expect to face?

Knowledge needed about

Speech:	acoustics and recognition
Words:	structure of words, their categories and meanings
Sentences:	structure of sentences and their meanings
Meaning/conceptualization:	sense disambiguation, Semantic representations, Translation equivalence,
Text/dialogue:	structure of texts or dialogs
Conventions:	cultural preferences and world knowledge, translation habbits

Traditional Tasks and components

Phonology:	from acoustic signal (speech) to words		
Morphology:	from words to morphemes (structure of words)		
Syntax:			
	word/morpheme categories	–	Part of Speech tagging
	sentence structure	–	syntactic analysis
Semantics:			
	word meaning	–	lexical semantics
	sentence meaning	–	compositional semantics
	word sense	–	words sense induction / disambigu
Pragmatics:	Language use, cultural conventions, world-knowledge		
Discourse:	How dialogs are structured, how text is structured		

What does NLP mean? how to build these components?

Goals of NLP?

Scientific: Build models of the human use of language and speech

Technological: Build models that serve in technological applications e.g. machine translation, speech systems, information extraction, etc.

Main NLP questions

1. **Models of language:** What are the kinds of things that people say and write?
2. **Language understanding:** What do these things mean?
3. **Effective algorithms:** for these questions

NLP questions and their interpretations

Question: Models of language

How to recognize (in formal terms) the things that people say and write?

Question: Language understanding

How to compute the meanings of these things?

(a) how to represent meaning?

(b) how to calculate the correct meaning?

Our aim: build models for the different NLP tasks

How to build NLP models?

Two views on modeling what people write and say:

Competence: Linguistic view on language processing
Excludes "non-linguistic factors", e.g. world knowledge, cultural preferences.
What people *should/could (in principle)* write and say?

Performance: Input-output view on language processing
Models of what people write and say

This course: *Performance* Models of NLP

Linguistic view: "Competence" Models

Formal language (set) theory as the tool:

A language	is	a set of word-sequences
A sentence	is	a sequence of words in the language
A grammar	is	a formal device defining the language
Grammaticality	is	a set membership test
Analyzing an utterance	is	assigning the correct structure

Is this view sufficient for NLP?

Why competence models not enough for processing (1)

Ambiguity: Many analyzes per utterance

Inherent: people tend to perceive very few analysis for an utterance, in most cases there is one preferred analysis:
"I saw (the man with the telescope)"
"I (saw (the man) (with the telescope))";
"I saw the dog with the telescope"

Technical: although many utterances are perceived by humans as *unambiguous*, linguistic grammars tend to assign very many possible analysis for each of these utterances!

The linguistic view of language as a set cannot resolve ambiguity.

Examples of ambiguity

Word-sense: words have different meanings in different contexts:
"west *bank* of the river" vs. "my savings in the *bank*"

Part-of-speech: word can have different categories:
"following" as "verb", "adj", or "noun".

Sentence structure: structure choice influences meaning
"The telegraphy and telephony services are important.",

Sentence meaning: semantics of sentence:
"She ran up a big bill" vs. *"She ran up a big hill"*

Word (spelling error): what is the correct word?
"I have been teading..."(teading$\in\{$*leading,reading,feeding*$\}$*)*

Why competence models not enough for processing (2)

Robustness: people can process "weird" utterances
"Who did Jo think said John saw him?"

Relative grammaticality: people see different levels of grammaticality:
"Those are the books you should read before talking about becomes difficult."
People disagree on how "grammatical" this utterance is.

Applications: in technological situations, language use is disturbed by hazard, resulting in ambiguous input (as in speech recognition, spelling correction etc.).

Some funny examples: Newspaper Headlines

- Iraqi Head Seeks Arms
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks

Examples from Chris Manning's website.

Summary of observations

Ambiguity: Humans disambiguate: competence models do not,

Robustness: Human language is not exactly a "set of utterances" in the formal sense: people process weird utterances while competence models do not,

Gradedness: Humans perceive graded grammaticality rather than absolute: competence models do not,

Other factors: Human processing involves extra-linguistic factors: competence models do not care about these factors,

Will any non-trivial NLP system ever be complete?

Performance: what should we model?

Given a task:

What to model? Input-output human language behavior (blackbox)

Assumptions: we assume the following:

- adult language use,
- a given domain of language use,

How to model? A function: input$\longrightarrow$output
The output that humans perceive as the most plausible for that input

Problem: how to resolve ambiguity, robustness, world-knowledge etc.?

Manifistations of Uncertainty

The problems of competence models are all manifistations of uncertainty!

Ambiguity:	uncertainty with respect to interpretation
Variability:	uncertainty in a specific realization for a semantic concept
Robustness:	uncertainty with respect to potential inputs
Lack of knowledge:	uncertainty!

We need: Theory of Uncertainty, Probability Theory

Sketch of probabilistic models

Functions: Inputs e.g. set of words+context / utterances,
Outputs e.g. set of POS-tags / syntactic analyzes,

Example: Part of Speech (PoS) tagging:

input	output		input	output
$input_1$	$output_1$		$\langle the, \underline{list} \rangle$	NN
$input_2$	$output_2$		$\langle We, \underline{list} \rangle$	VB
$\vdots$	$\vdots$		$\vdots$	$\vdots$

Model: A probability function over input–output pairs

$$P : Inputs \times Outputs \longrightarrow [0, 1]$$

What does this mean?

A language is a probability function over the set of all allowed input–output pairs

- What qualities does this function have?
- How to obtain this function?
- How to use this function?

Plausibility: kind of relative (as opposed to absolute) "grammaticality" :

> The value $P(\langle i, o \rangle)$ expresses the intuitive question: "How plausible is the pair $\langle i, o \rangle$ in language use?"

Does this solve the manifistations of uncertainty?

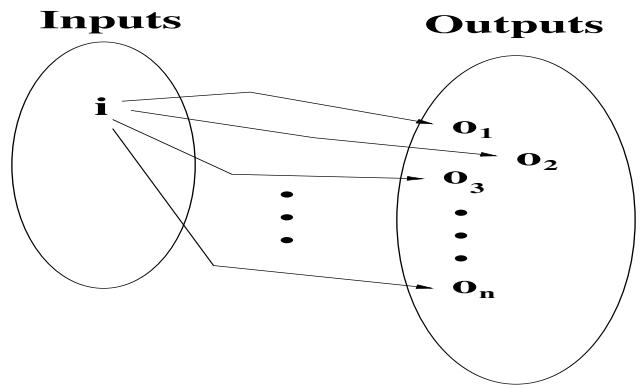
Example: ambiguity resolution (disambiguation)

Given $P : \mathit{Inputs} \times \mathit{Outputs} \longrightarrow [0, 1]$

Suppose $o_1, \cdots, o_n$ are possible outputs for input i

Q: how to select the preferred output o^* for input i?

A: select $o^* = \arg\max_{o \in \{o_1, \cdots, o_n\}} P(\langle i, o \rangle)$



$\arg\max_{a \in A} P(a)$: select that a from A for which $P(a)$ is maximal

Questions and Issues

Questions	**Answers**
How to define set of input-output pairs?	Use formal grammars
How to define the probabilities $P(i, o)$?	Probabilistic grammars
How to obtain these probabilities?	Statistics from corpora
What (efficient) algorithms we need?	Technical solutions
How to measure success of a model?	Evaluation methods

NLP becomes Probabilistic, Data-Driven and Empirical

Related issues: Learning and Adaptive behavior

Example Probabilistic Model

Input-Output: Utterance $\Longrightarrow$ Parse

Input-Output Sets: A Context-Free Grammar (CFG)
Inputs = String language (set of sentences)
Outputs=Tree language (set of parses)

Probabilistic version: A Probabilistic CFG defining how to calculate for every sentence U and parse T: $P(\langle U, T \rangle)$

Parsing: Computing a tree with the highest probability t^* for a sentence s

$$t^* = \arg \max_{t \in G(s)} P(\langle s, t \rangle)$$

Where $G(s)$ is set of possible translations.

Structure. . .Probability. . .Statistics. . .Evaluation

For every task we will discuss the following:

- **Structure:** What linguistic knowledge specifies the input-output relation?
- **Probability:** How to define the model probabilities over input-output pairs?
- **Estimation:** How to estimate/learn the model probabilities from data?
- **Evaluation:** How to evaluate the resulting model on test data?

Structure of First Part of Course

1. Introduction
2. Issues

Math:	Elementary Probability Theory
Data:	Statistics, Estimation and Learning
Models:	Language Models over Linguistic Structure
Smoothing:	Sparse-Data and Smoothing Methods
Evaluation:	Evaluation Measures: Empirical and Information Theoretic

Probabilistic Models over Words, POS tags and Parse-Trees

II. Language Models

Language Models (broad sense)

Language Models (broad sense)

Computational models of human language.

- models that are detailed enough to be able to **write computer programs** to perform various tasks involving spoken/written natural language

Language Models (broad sense)

Computational models of human language.

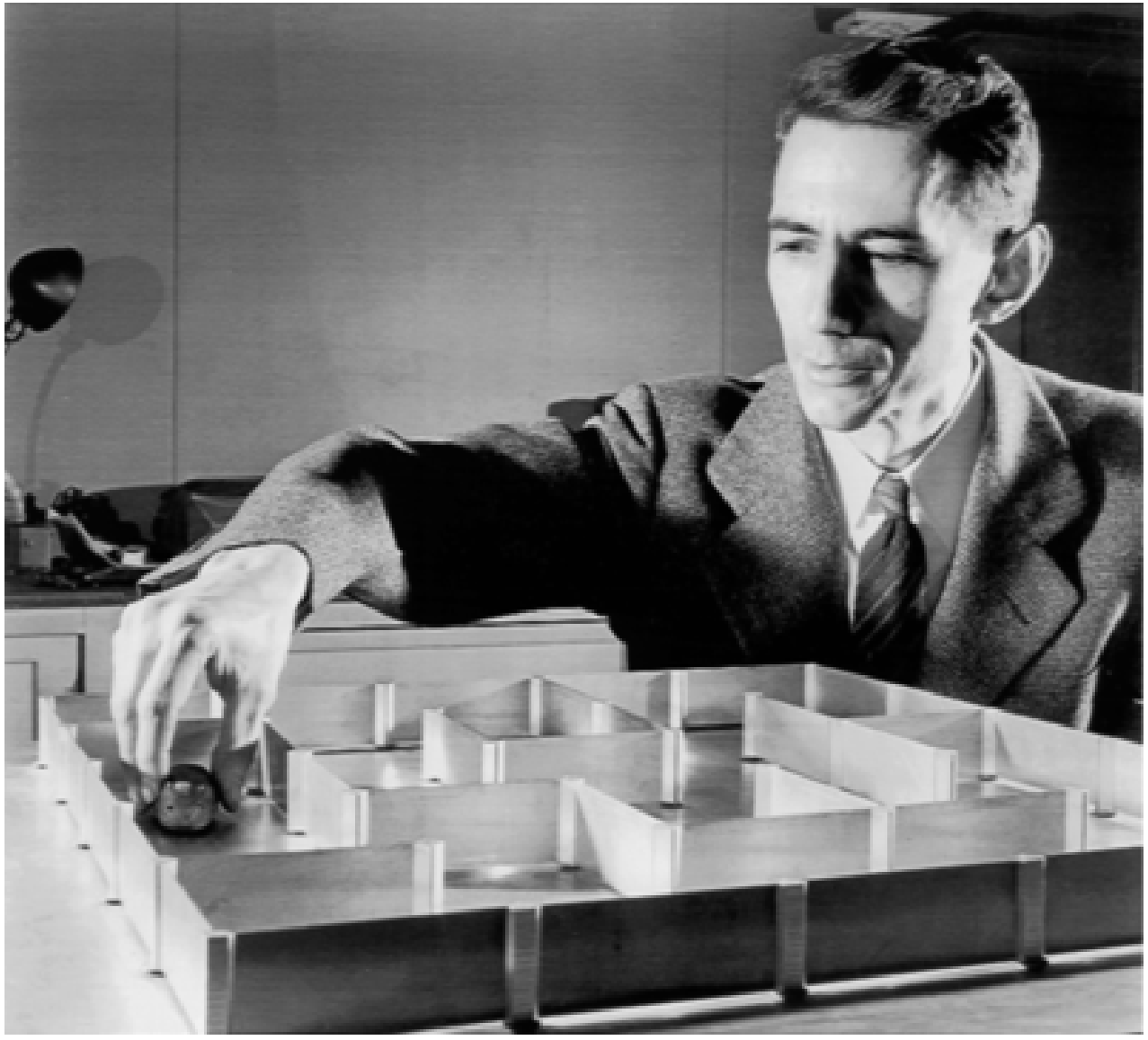
- models that are detailed enough to be able to **write computer programs** to perform various tasks involving spoken/written natural language

Scientific Goal: Build models of the human use of language and speech (computational linguistics)

Technological Goal: Build models that serve in technological applications e.g. machine translation, speech systems, information extraction, etc. (Natural Language Processing)

Outline

Claude Shannon, 1916-2001



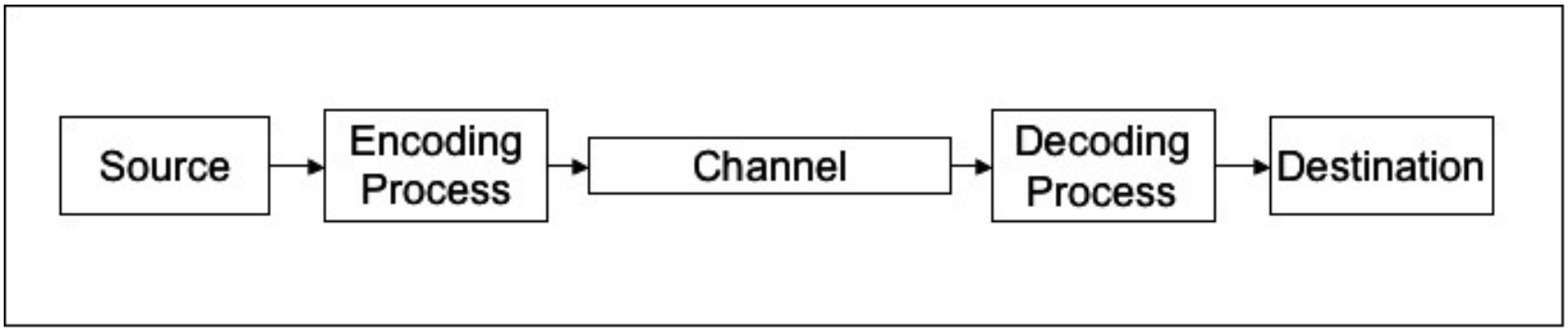
Shannon with his self-learning mouse Theseus

Shannon's information theory

(Weaver, 1949)

- Three levels of analysis, Shannon deals with the first only:
  - **Technical level:** How accurately can the symbols of communication be transmitted?
  - **Semantic level:** How precisely do the transmitted symbols convey the desired meaning?
  - **Effectiveness level:** How effectively does the received meaning affect conduct in the desired way?
- At the technical level, the content of communicative act is irrelevant; the source is viewed as a *stochastic process*;
- **Shannon's concept of information**: reduction in uncertainty about the source;

Shannon's noisy channel model



Shannon's questions

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

Shannon's questions

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?

Shannon's questions

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?
2. What is the best way to encode messages from the source?

Shannon's questions

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?
2. What is the best way to encode messages from the source?
3. What is the best way to measure the amount of information from a source?

Shannon's questions

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?
2. What is the best way to encode messages from the source?
3. What is the best way to measure the amount of information from a source?
4. What is the best guess on the source's intended message if the signal has been distorted by noise?

Question 1: Encoding messages

What is the best way to encode messages from the source?

Question 1: Encoding messages

What is the best way to encode messages from the source?

- Frequent messages should be assigned a short encoding;

Question 1: Encoding messages

What is the best way to encode messages from the source?

- Frequent messages should be assigned a short encoding;
- Highly predictable sources require less channel capacity;

Question 2: Measuring information (1)

Imagine we have a distribution $p(x)$ over potential messages $x \in X$

- Information $I(x)$ contained in a message is a measure of surprisal
- If a message x is known (deterministic): $I(x) = 0$
- Less likely the message - more information in the message
- **Additivity:** information in 2 independent msgs = sum of informations in the msgs

Self information for a message x: $I(x) = -\log_2 p(x)$

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

- Another view: desiderata for a measure H of information:
  - H should be *continuous*

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

- Another view: desiderata for a measure H of information:
  - H should be *continuous*
  - If all $p(x)$ are equal, H should *increase monotonously* with $|X|$

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

- Another view: desiderata for a measure H of information:
  - H should be *continuous*
  - If all $p(x)$ are equal, H should *increase monotonously* with $|X|$
  - If a choice can be broken down, H should be the weighted sum of the H's of successive choices (*additivity*).

Question 2: Measuring information (2)

One view: the entropy is an average self-information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

- Another view: desiderata for a measure H of information:
  - H should be *continuous*
  - If all $p(x)$ are equal, H should *increase monotonously* with $|X|$
  - If a choice can be broken down, H should be the weighted sum of the H's of successive choices (*additivity*).

Information theory

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

Information theory

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?

Information theory

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?
2. What is the best way to encode messages from the source?

Information theory

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?
2. What is the best way to encode messages from the source?
   - Frequent messages should be assigned a short encoding;
   - Highly predictable sources require less channel capacity;

Information theory

Shannon (1948)

- Source $\rightarrow$ Encoder $\rightarrow$ Channel $\rightarrow$ Decoder $\rightarrow$ Destination

1. How much information can go in principle through the channel per unit of time?
2. What is the best way to encode messages from the source?
   - Frequent messages should be assigned a short encoding;
   - Highly predictable sources require less channel capacity;
3. What is the best way to measure the amount of information H) from a source?

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Noisy Channel Model in NLP

- Forget about technical answers to the questions

Noisy Channel Model in NLP

- Forget about technical answers to the questions
- Noisy Channel becomes a useful metaphor. E.g., machine translation: think of French as very noisy English.

Noisy Channel Model in NLP

- Forget about technical answers to the questions
- Noisy Channel becomes a useful metaphor. E.g., machine translation: think of French as very noisy English.
- Question 4: What is the best guess on the source's intended message if the signal has been distorted by noise?

Noisy Channel Model in NLP

- Forget about technical answers to the questions
- Noisy Channel becomes a useful metaphor. E.g., machine translation: think of French as very noisy English.
- Question 4: What is the best guess on the source's intended message if the signal has been distorted by noise?
- Source model $\times$ Distortion model

$$\text{argmax}_e P(e)P(f|e)$$

(e is message or English, f is signal or French)

Source models: ngrams

f $P(f)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$
f, a, g	$P(g)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$
f, a, g	$P(g)$
f, a, g, g	$P(g)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$
f, a, g	$P(g)$
f, a, g, g	$P(g)$
f, a, g, g, b	$P(b)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$
f, a, g	$P(g)$
f, a, g, g	$P(g)$
f, a, g, g, b	$P(b)$
f, a, g, g, b, a	$P(a)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$
f, a, g	$P(g)$
f, a, g, g	$P(g)$
f, a, g, g, b	$P(b)$
f, a, g, g, b, a	$P(a)$
f, a, g, g, b, a, c	$P(c)$

Source models: ngrams

f	$P(f)$
f, a	$P(a)$
f, a, g	$P(g)$
f, a, g, g	$P(g)$
f, a, g, g, b	$P(b)$
f, a, g, g, b, a	$P(a)$
f, a, g, g, b, a, c	$P(c)$

Markov-order: $0 =$ unigram model.

Source models: ngrams

f $P(f|\#)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$
f, a, g	$P(g\|a)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$
f, a, g	$P(g\|a)$
f, a, g, g	$P(g\|g)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$
f, a, g	$P(g\|a)$
f, a, g, g	$P(g\|g)$
f, a, g, g, b	$P(b\|g)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$
f, a, g	$P(g\|a)$
f, a, g, g	$P(g\|g)$
f, a, g, g, b	$P(b\|g)$
f, a, g, g, b, a	$P(a\|b)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$
f, a, g	$P(g\|a)$
f, a, g, g	$P(g\|g)$
f, a, g, g, b	$P(b\|g)$
f, a, g, g, b, a	$P(a\|b)$
f, a, g, g, b, a, c	$P(c\|a)$

Source models: ngrams

f	$P(f\|\#)$
f, a	$P(a\|f)$
f, a, g	$P(g\|a)$
f, a, g, g	$P(g\|g)$
f, a, g, g, b	$P(b\|g)$
f, a, g, g, b, a	$P(a\|b)$
f, a, g, g, b, a, c	$P(c\|a)$

Markov-order: $1 =$ bigram model.

Source models: ngrams

f $P(f|\#\#)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$
f, a, g, g	$P(g\|ag)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$
f, a, g, g	$P(g\|ag)$
f, a, g, g, b	$P(b\|gg)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$
f, a, g, g	$P(g\|ag)$
f, a, g, g, b	$P(b\|gg)$
f, a, g, g, b, a	$P(a\|gb)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$
f, a, g, g	$P(g\|ag)$
f, a, g, g, b	$P(b\|gg)$
f, a, g, g, b, a	$P(a\|gb)$
f, a, g, g, b, a, c	$P(c\|ba)$

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$
f, a, g, g	$P(g\|ag)$
f, a, g, g, b	$P(b\|gg)$
f, a, g, g, b, a	$P(a\|gb)$
f, a, g, g, b, a, c	$P(c\|ba)$

Markov-order: $2 =$ trigram model.

Source models: ngrams

f	$P(f\|\#\#)$
f, a	$P(a\|\#f)$
f, a, g	$P(g\|fa)$
f, a, g, g	$P(g\|ag)$
f, a, g, g, b	$P(b\|gg)$
f, a, g, g, b, a	$P(a\|gb)$
f, a, g, g, b, a, c	$P(c\|ba)$

Markov-order: 2 = trigram model.
We can easily generalize to Markov-order n-1 = ngram model.

Source models: ngrams on characters

Approximations of English based on character transition probabilities:

0-order: XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

1st-order: OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

2nd-order: ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

3d-order: IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Source models: ngrams on words

Approximations of English based on word transition probabilities:

1st-order: REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE

2nd-order: THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Shannon's influence

- Shannon's 1948 paper has been extraordinary influential in many fields.
- In **language modelling**, his application of ngram-models was instantly popular. Still widely used as language models (in the narrowest sense): to assign probabilities to sequences.
- Provoked Noam Chomsky to dispute the inadequacy of Markov models for describing syntactic structure.
- Established the need for probabilistic models of language (although Chomsky made them impopular in linguistics for a while).
- A language model (= source model) in combination with a task model (= distortion model) provides a useful division of labor in many NLP tasks (next weeks).

IV. Probability Theory Reminder

Probability Theory Reminder

Deals with averages of mass phenomena.

Experiments, Events and Probabilities (1)

Experimental model: a setting with a set of possible "outcomes".
Example: Casting a die with any of the six sides as outcomes.

Trial: A single execution of the experiment.
Example: Casting the die once with an outcome.

Sample space (Ω): Set of mutually exclusive outcomes of the experiment
Example: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Event: Any subset of the sample space Ω.
Example: $\{X \in \Omega | even(X)\}$ denotes the set of all even outcomes in Ω.

Probability Model

A Probability Model is a pair $< E, P >$ where $E = 2^{\Omega}$ is the set of all events over Ω, and P is a probability mass function which fulfills the following properties:

[0,1] Range: $P : E \rightarrow [0, 1]$

Certain event: $P(\Omega) = 1$

Mutually exclusive events

$$\forall A, B \in E : \quad (A \cap B = \emptyset) \quad \Longrightarrow \quad P(A \cup B) = P(A) + P(B)$$

Independence and Probability

Joint Probability $(P(A \cap B))$

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

Conditional probability (A given B):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}$$

Definition: A and B are called *independent* iff

$$P(A \cap B) = P(A)P(B)$$

Corollaries: A and B independent iff

$$P(A) = P(A|B), \quad P(B) = P(B|A)$$

Conditional Probability: Review

Example of independence

In the Netherlands it rains 30 % of the days but the weather is changeable. We can guess the weather today is independent of the weather yesterday:

$$P(\text{rain today}|\text{rain yesterday}) = P(\text{rain today})$$

$$P(\textit{rain today}, \textit{rain yest}) = P(\textit{rain today})P(\textit{rain yest}.)$$

Not independent

In the Sahara rain is very rare but if it rains it rains several days in succession. Here $P(\textit{rain today})$ is small but
$P(\textit{rain today}|\textit{rain yesterday})$ is great.

Next time

- N-grams for language modelling
- Noisy Channel model
- Detecting non-words, spelling correction