Natural Language Models & Interfaces, 2012

Jelle Zuidema

ILLC, Universiteit van Amsterdam
(slides in part adapted from Tejaswini Deoskar, Yoav Seginer and Khalil Sima'an)

4 April 2012

I. What is language?

Symbolicism

- Word meanings are conventional and arbitrary: even e.g. onomatopoeia: cock-a-doodle-do, cocorico, kukeleku
- Adult speakers know $\approx$ 10.000 words; children learn these words over the course of just a few years.
- Languages are transmitted culturally, and slowly change over the course of a number of generations, giving rise to an enormous variety of over $>$ 6000 languages.
- Every language has its own vocabulary, and languages differ in how conceptual structure get expressed in words (e.g. compare English-Dutch, pot vs. pan, on the wall vs. on the table)

Combinatorial Phonology

- Words are built-up from (meaningless) basic speech sounds, the phonemes;
- Phonemes are defined as the minimal difference in sound that corresponds to a difference in meaning. E.g. minimal pairs:

$$\left| \begin{array}{cc} bed & bad \\ bet & bat \end{array} \right| \mapsto /e/, /a/, /d/, /t/$$

- Phonemes are different in every language (and dialect), but phonemic coding is universal.
- 3 vowels (/i/,/u/,/a/) in Greenlandic, to 12 (English) or even 15 (Norwegian)

Compositional Semantics

- Sentences are built-up from meaningful words.
- Words can be built from meaningful morphemes. E.g.: "he walk-s", disproof, disallow, rearrange
- The meaning of a larger whole is determined by the meaning of the words and the way they are put together (compositionality), i.e. word order and morphological marking and agreement: e.g. "In vino veritas est."

- E.g. Vietnamese (an "isolating language"):

Khi tôidên nhà ban tôi,chúng tôibắtdâulàmbài.
whenI comehousefriendI PLURALI begin do lesson
"When I came to my friend's house, we began to do lessons."

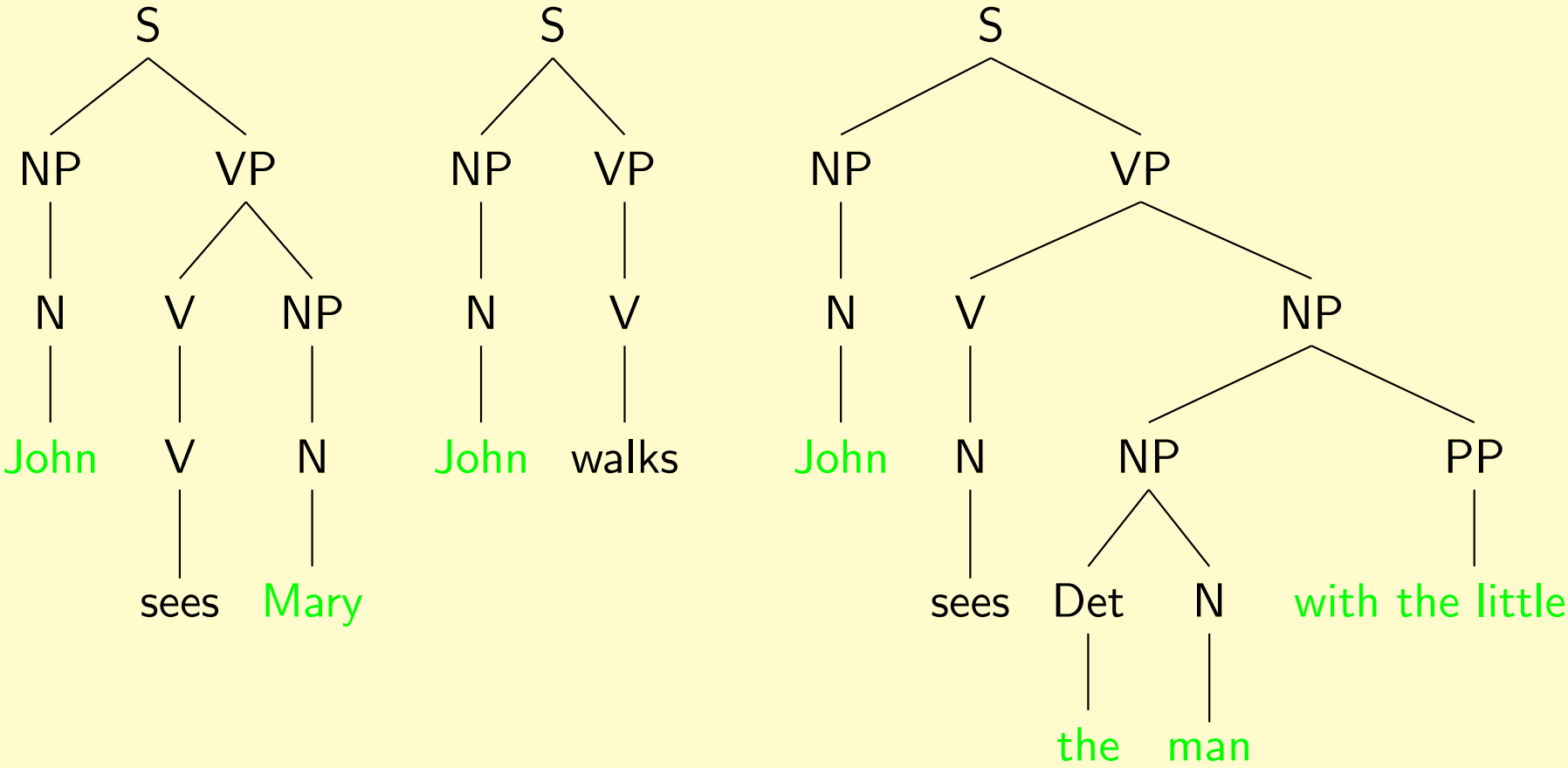
- E.g. Chukchi (Siberia; a "polysynthetic language"):

tϵ- meyηϵ-levtϵ-pϵγt- ϵrkϵn
greathead ache 1stSINGULARIMPERFECT
"I have a fierce head-ache"

(Examples from Comrie, 1981)

Recursive Syntax

- Sentences have a constituent structure. E.g.:



- Constituents of a certain type, can contain constituents of the same type (recursion), "infinite use of finite means".

1. Gilligan claims that Kelly saw Campbell help Blair deceive the public. (TAIL RECURSION)
2. Gilligan behaupte dass Kelly Campbell Blair das Publikum belügen helfen sah. (CENTER EMBEDDING)
3. Gilligan beweert dat Kelly Campbell Blair het publiek zag helpen bedriegen. (MILD CONTEXT-SENSITIVITY)

II. Language Models

Language Models (broad sense)

Language Models (broad sense)

The art and science of creating *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)

The art and science of creating *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)

Examples of natural language processing

A program that

- Performs spelling and grammar checking/correction in documents
- Gives or understands commands (*Turn the TV on*)
- Can find answers to your questions on the internet *Who wrote the Harry Potter books?*
- Dialogue systems *Booking a flight, getting directions to a destination*
- Translating text/speech from one language to another.

An example of an Internet Question-Answering System

A person writes a question: `Who is <NAME >?`

Target:	the computer must search for relevant texts
Understand	that the text is about the person in question (Victoria Beckham=Posh Spice/ President=Mr. President=Mr. Obama=The president of the United States).
Action:	search and print out answers to the reply.

Understand: which texts are typical of information about individuals such as: Birth date/place (was born in), education (he graduated from), work (he was chief executive of) Etc. ..)

Information Extraction

Text

```
Bogota, 9 Jan- Richardo Alfonso Castellar, Mayor of Achi,
In the Northern department of Boliva, who was kidnapped on
5 January, apparently by Army of National Liberation (ELN)
Guerrilas, was found dead today, according to authorities.
Castellar was kidnapped on ....
```

Summary

Date	05 Jan 90
Location	Columbia: Bolivar (Department): Achi (Town)
Type	Kidnapping
Weapon	*
Victim	Ricardo Alfonso Castellar
...	

- Read a text
- Derive assertions that can be put into a structured database.

Example from Claire Cardie's CS472 class

Example dialogue 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 habits

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	–	translation equivalents
Pragmatics:	Language use, cultural conventions, world-knowledge		
Discourse:	How dialogs are structured, how text is structured		

How to build these components?

Why is NLP hard?

- Would be easy if Natural Language were like programming languages: unambiguous, fully specified, fully observable...
- ... but Natural Languages are:
  - **Massively** ambiguous, at **all** levels of analysis.
  - still badly understood; which rules do humans use?
  - largely unobservable; what are the intermediate representations?
- NLP is thought to be **AI-complete**/ AI-hard (difficulty is equivalent to solving the central AI problem - making computers as intelligent as people)

Ambiguity in human languages

Speech recognition

What is the correct utterance:
"I scream" vs. "Ice-cream"

- "It's very hard to recognize speech"
- "It's very easy to wreak a nice beach."

Word (spelling error):

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

Part-of-speech

Word can have different categories:
Can you open this can?

Ambiguity in human languages

Word-sense

Words have different meanings in different contexts: "west *bank* of the river" vs. "my savings in the *bank*"

Sentence structure

Structure-choice influences meaning
"I saw the man with the telescope"
"I saw John and Bill's sister going to the cinema."

Sentence meaning:

Semantics of sentence:
"She ran up a big bill" vs. *"She ran up a big hill"*

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.

NLP tasks & uncertainty

NLP tasks like Speech Recognition, Parsing, Natural Language Generation, Natural Language Understanding, Information Extraction....

NLP tasks & uncertainty

NLP tasks like Speech Recognition, Parsing, Natural Language Generation, Natural Language Understanding, Information Extraction....

- always involve a mapping from a complex representation (e.g. spoken form) to a complex representation (e.g., written form)
- always involve uncertainty due to inherent ambiguity and our limited knowledge.

NLP tasks & uncertainty

NLP tasks like Speech Recognition, Parsing, Natural Language Generation, Natural Language Understanding, Information Extraction....

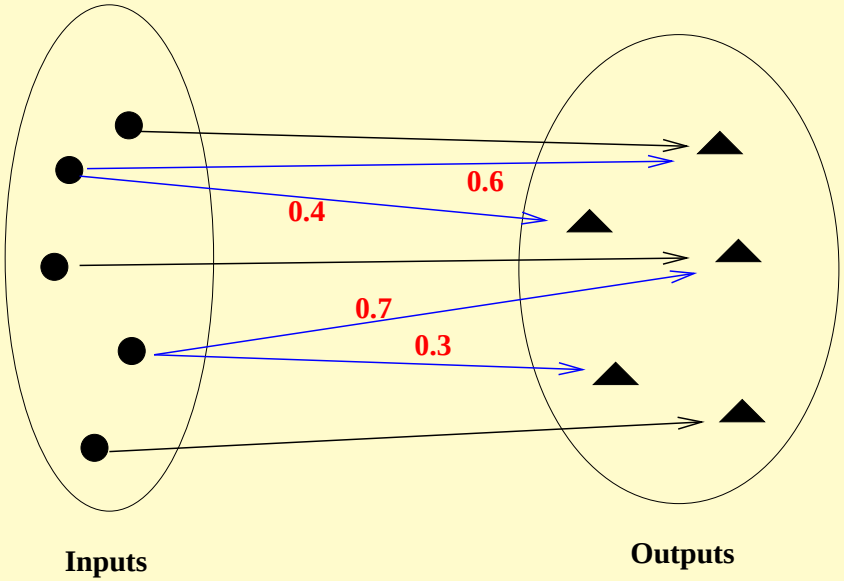
- always involve a mapping from a complex representation (e.g. spoken form) to a complex representation (e.g., written form)
- always involve uncertainty due to inherent ambiguity and our limited knowledge.

We need: A Theory of Uncertainty: Probability Theory

Extention to probabilistic/statistical model

Given several options, we output the most likely outcome.

spellings	output	prob.	rank
teaching	I have been teaching a course	0.49	1
leading	I have been leading a course	0.15	3
feeling	I have been feeling a course	0.12	4



How can we do that?

Language Models (narrower sense)

A language model provides an estimate of the chances of various language items (words, word combinations, phrases, etc.)

A language model is used to determine the most likely possibility, given by an NLP algorithm generating say spelling corrections,POS-tags, translations, etc.

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

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

Today's lecture

What is language?

symbolicism, phonology, compositionality, recursion

very useful, very complex, very diverse

Language Models

broad sense: formal models of language processing and learning

narrower sense: assigning probabilities to linguistic structure

Administrative issues

organization, requirements, grading

Information & Probability Theory

Little digression: Claude Shannon

Noisy Channel Model

Language Models in the narrowest sense

Ngrams

Probability Theory Reminder

III. Administrative issues

Administrative Issues

Lecturer: Jelle Zuidema (ILLC)

Assistants: Maarten van de Velden and Frank Smit

Textbook: D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2008, 2nd edition (1st edition at your own risk).

Blackboard: Contact TA's if you aren't enrolled.

Your tasks (per week):

1. Attend lectures, 4hrs
2. Participate in computer labs, 4hrs
3. Finish assignments and write papers (4hrs)
4. Review notes & read assigned textbook chapters (4hrs)
5. Prepare presentation (1hr *8)

Grading/Evaluation

- Exam 1 (25% of grade)
- Exam 2 (25% of grade)
- 6 assignments (pass/fail; 3% of grade each)
- Presentation & model code (5% of grade)
- First version paper complete & on time (pass/fail; 3% of grade each)
- Final paper (21% of grade)

Outline

What is language?

- symbolicism, phonology, compositionality, recursion
- very useful, very complex, very diverse

Language Models

- broad sense: formal models of language processing and learning
- narrower sense: assigning probabilities to linguistic structure

Administrative issues

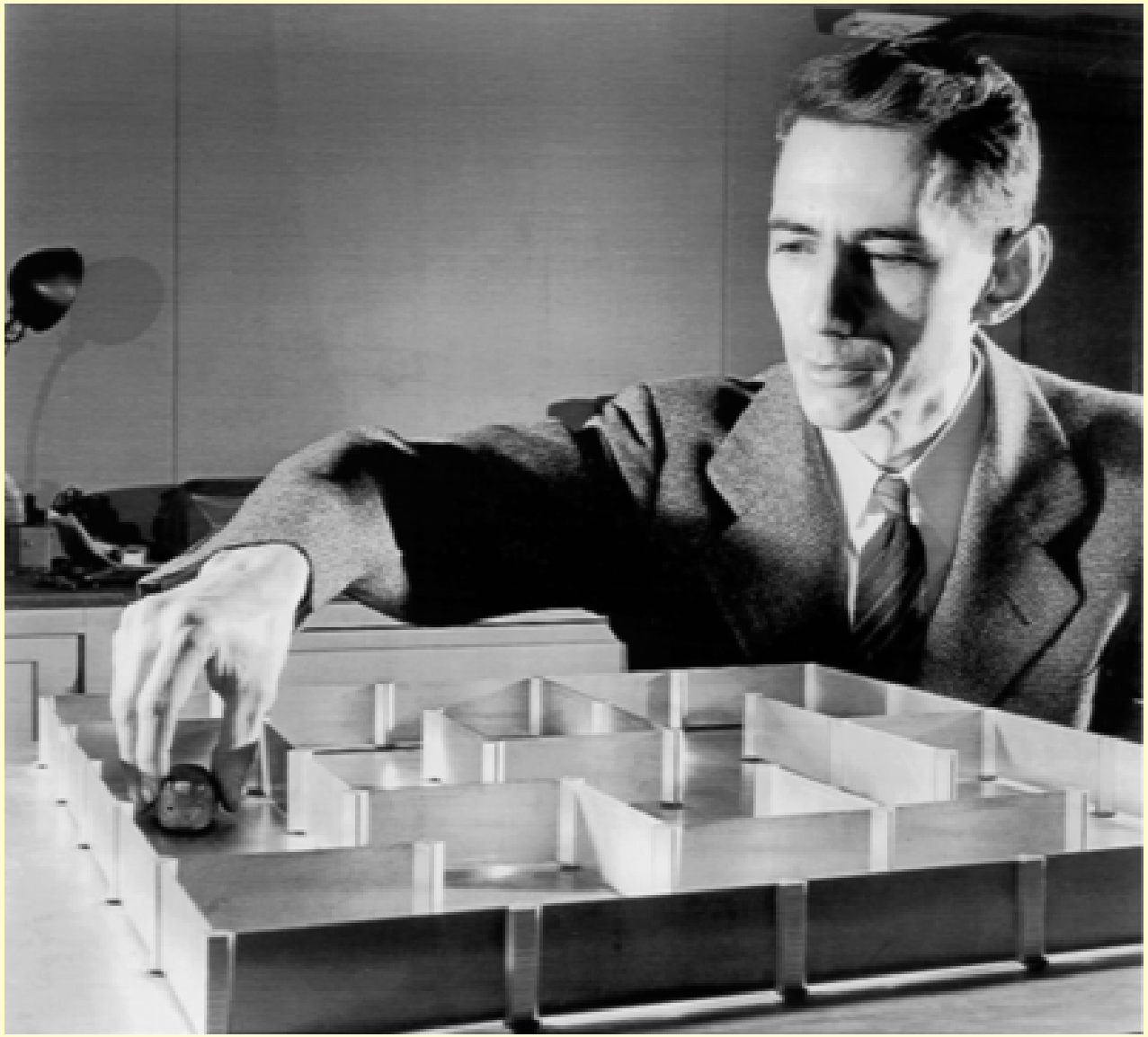
- organization, requirements, grading

Information & Probability Theory

- Little digression: Claude Shannon
- Noisy Channel Model
- Language Models in the narrowest sense
- Ngrams

Probability Theory Reminder

Claude Shannon, 1916-2001



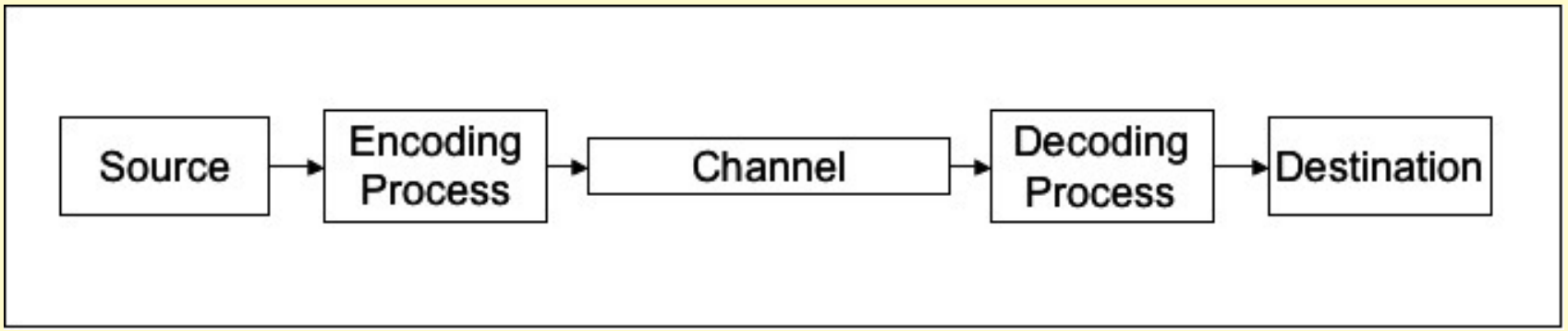
(Shannon with his selflearning mouse Theseus; he also built rocket-powered flying disks, Ultimate Machine, ...)

Shannon's information theory

(Weaver, 1949)

- Three levels of analysis, Shannon deals with the first only:
  - Technical level
  - Semantic level
  - Effectiveness level
- 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;
- (Note: a subjectivist interpretation of probabilities)

Shannon's noisy channel model



Shannon's questions

Shannon (1948)

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

Shannon's questions

Shannon (1948)

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

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

Shannon's questions

Shannon (1948)

- Source $\longrightarrow$ Encoder $\longrightarrow$ Channel $\longrightarrow$ Decoder $\longrightarrow$ 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 $\longrightarrow$ Encoder $\longrightarrow$ Channel $\longrightarrow$ Decoder $\longrightarrow$ 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 $H(p_1, p_2, \ldots, p_n)$ from a source $p_1, p_2, \ldots, p_n$?

Shannon's questions

Shannon (1948)

- Source $\longrightarrow$ Encoder $\longrightarrow$ Channel $\longrightarrow$ Decoder $\longrightarrow$ 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 $H(p_1, p_2, \ldots, p_n)$ from a source $p_1, p_2, \ldots, p_n$?
4. What is the best guess on the source's intended message if the signal has been distorted by noise?

Question 1: Measuring channel capacity

- If time and channel are discrete, the capacity is simply the maximum number of bits per units of time...

Question 1: Measuring channel capacity

- If time and channel are discrete, the capacity is simply the maximum number of bits per units of time...
- ... but what if time and channel are continuous?

$$C = \lim_{T \to \infty} \frac{\log_2 N(T)}{T} \tag{1}$$

Question 2: Encoding messages

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

Question 2: Encoding messages

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

- Frequent messages should be assigned a short encoding;

Question 2: 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 3: Measuring information

Question 3: Measuring information

- If all messages are equally probable, we can count the number of different possible messages...

Question 3: Measuring information

- If all messages are equally probable, we can count the number of different possible messages...
- ... but what if they are not equally probable?

Question 3: Measuring information

- If all messages are equally probable, we can count the number of different possible messages...
- ... but what if they are not equally probable?
- Desiderata for a measure H of information:
  - H should be *continuous*

Question 3: Measuring information

- If all messages are equally probable, we can count the number of different possible messages...
- ... but what if they are not equally probable?
- Desiderata for a measure H of information:
  - H should be *continuous*
  - If all p_i are equal, H should *increase monotonously* with n

Question 3: Measuring information

- If all messages are equally probable, we can count the number of different possible messages...
- ... but what if they are not equally probable?
- Desiderata for a measure H of information:
  - H should be *continuous*
  - If all p_i are equal, H should *increase monotonously* with n
  - If a choice can be broken down, H should be the weighted sum of the H's of successive choices (*additivity*).

Question 3: Measuring information

- If all messages are equally probable, we can count the number of different possible messages...
- ... but what if they are not equally probable?
- Desiderata for a measure H of information:
  - H should be *continuous*
  - If all p_i are equal, H should *increase monotonously* with n
  - If a choice can be broken down, H should be the weighted sum of the H's of successive choices (*additivity*).
- $$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{2}$$

Information theory

Shannon (1948)

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

Information theory

Shannon (1948)

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

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

$$C = \lim_{T \to \infty} \frac{\log_2 N(T)}{T}$$

Information theory

Shannon (1948)

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

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

$$C = \lim_{T \to \infty} \frac{\log_2 N(T)}{T}$$

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?

$$C = \lim_{T \to \infty} \frac{\log_2 N(T)}{T}$$

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?

$$C = \lim_{T \to \infty} \frac{\log_2 N(T)}{T}$$

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 $H(p_1, p_2, \ldots, p_n)$ from a source $p_1, p_2, \ldots, p_n$?

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

Noisy Channel Model in NLP

- Forget about technical answers to questions 1-3...

Noisy Channel Model in NLP

- Forget about technical answers to questions 1-3...
- 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 questions 1-3...
- 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 questions 1-3...
- 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 demonstrate the inadequacy of Markov models for describing syntactic structure (week 3).
- 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 (week 2).

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 is an event 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 week

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

Assignment

Due before next Wednesday

1. Assignment 1 (on Blackboard)

Reading: Jurafsky and Martin Chapter 1, 2.1 and 4.1-4.4.