Statistical Methods in Natural Language Processing

1. Introduction, Probability

Pavel Pecina, Jindřich Helcl

7 October, 2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Lecturers

Pavel Pecina - lectures, practicals
ÚFAL MFF UK, Room S422
pecina@ufal.mff.cuni.cz

Jindřich Helcl – homework assignments, exam
ÚFAL MFF UK, office N233
helcl@ufal.mff.cuni.cz

Course Logistics

- Webpage: https://ufal.cz/courses/npfl147
- Lectures on Tuesdays @ 12:20, Room **S9**
- Practicals on Tuesdays @ 14:00, Room **S1** (Moodle Quizzes, Q&A's)
- First lecture **Oct 7**
- No lecture/practicals **Oct 28** (national holiday)
- No lecture/practicals **Nov 25**
- Homework projects assigned during the semester
- Exam date (probable) **Jan 13, 2026**

Homework Assignments

- **Three homework assignments** with fixed deadlines
- To be worked on independently
- Require a substantial amount of programming/experimentation/reporting
- The assignments will be awarded by 0–100 points each
- Late submissions up to 2 weeks → 50% point reduction
- Submissions received later than 2 weeks → 0 points

Exam

- **Open-book written test**
- The maximum duration of the test is 90 minutes.
- The test will be graded by 0–100 points.

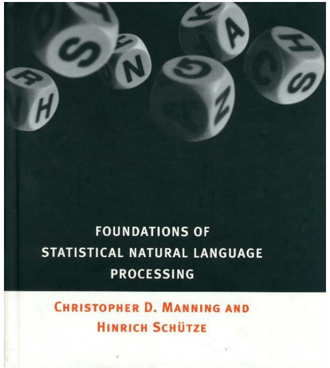
Passing Requirements

- Completion of both the homework assignments and exam is required
- Students need to earn **at least 50 points for each assignment** (before late submission penalization) and **at least 50 points for the test.**
- The points received for the assignments and test will be available in **SIS**.
- The **final grade** will be based on the average results of the exam test and the three homework assignments, **all four weighted equally**.

Readings

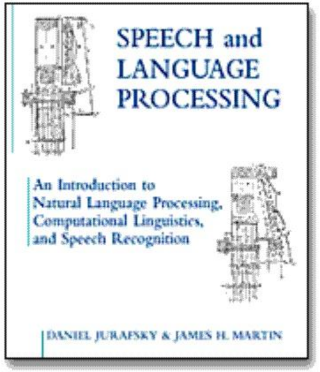
Foundations of Statistical Natural Language Processing

Manning, C. D. and H. Schütze. *MIT Press*. 1999. ISBN 0-262-13360-1.



Speech and Language Processing

Jurafsky, D. and J. H. Martin. *Prentice-Hall*. 2000. ISBN 0-13-095069-6



Course Segments

1. Introduction, probability, essential information theory
2. Statistical language modelling (n-gram)
3. Statistical properties of words
4. Word embeddings
5. Hidden Markov models, Tagging

Summary

- The course materials will be available on the course webpage.
- If you have questions, drop us a line.

Introduction

Why is NLP difficult?

- many “words”, many “phenomena” → many “rules”
  - **OED: 400k words; Finnish lexicon (of forms): ~2 . 107**
  - sentences, clauses, phrases, constituents, coordination, negation, imperatives/questions, inflections, parts of speech, pronunciation, topic/focus, and much more!
- irregularity (exceptions, exceptions to the exceptions, ...)
  - plural forms
    - **potato → potato es (tomato, hero,...); photo → photo s**
    - and even: **both mango -> mango s   or → mango es**
  - Adjective / Noun order
    - **new book, electrical engineering, general regulations, flower garden, garden flower**
    - but **Governor General**

Other difficulties in NLP

Ambiguity

- **books**
  - NOUN or VERB**?**
  - *you need many books* vs. *she books her flights online*
- **No left turn weekdays 4-6 pm / except transit vehicles**
  - when may transit vehicles turn: Always? Never?
- **Thank you for not smoking, drinking, eating or playing radios without earphones.**
  - Thank you for not eating without earphones??
  - or even: Thank you for not drinking without earphones!?
- **My neighbor's hat was taken by wind. He tried to catch it.**
  - ...catch the wind or ...catch the hat ?

(Categorical) Rules or Statistics?

Preferences:

- clear cases: context clues: she books → books is a verb
  - rule: if an ambiguous word (verb/nonverb) is preceded by a matching personal pronoun → word is a verb
- less clear cases: pronoun reference
  - she/he/it refers to the most recent noun or pronoun (?) (but maybe we can specify exceptions)
- selectional:
  - catching hat >> catching wind (but why not?)
- semantic:
  - never thank for drinking in a bus! (but what about the earphones?)

Solutions

- Don’t guess if you know:
  - morphology (inflections)
  - lexicons (lists of words)
  - unambiguous names
  - perhaps some (really) fixed phrases
  - syntactic rules?

- Use **statistics** (based on real-world data!) for preferences

- (Combination also possible)

Statistical NLP

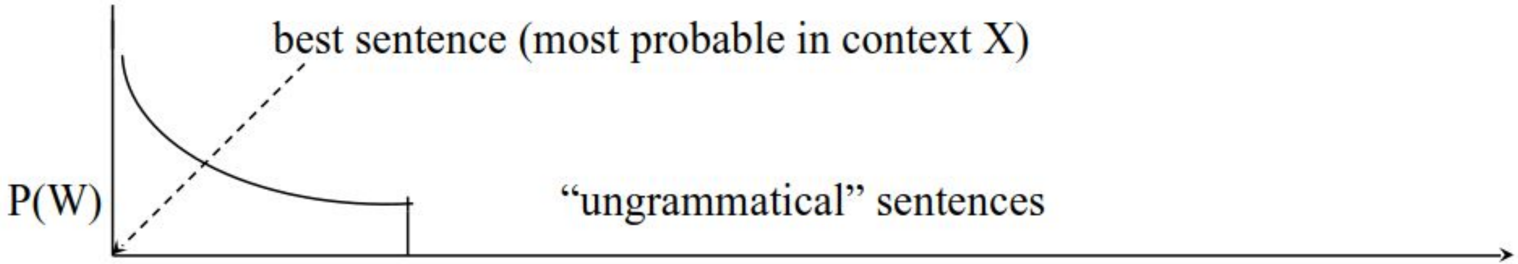
Imagine:

- Each sentence $W = \{ w_1, w_2, ..., w_n\}$ gets a probability $P(W|X)$ in a context X (think of it in the intuitive sense for now)
- For every possible context X, sort all the imaginable sentences W according to $P(W|X)$:
- Ideal situation:

Statistical NLP

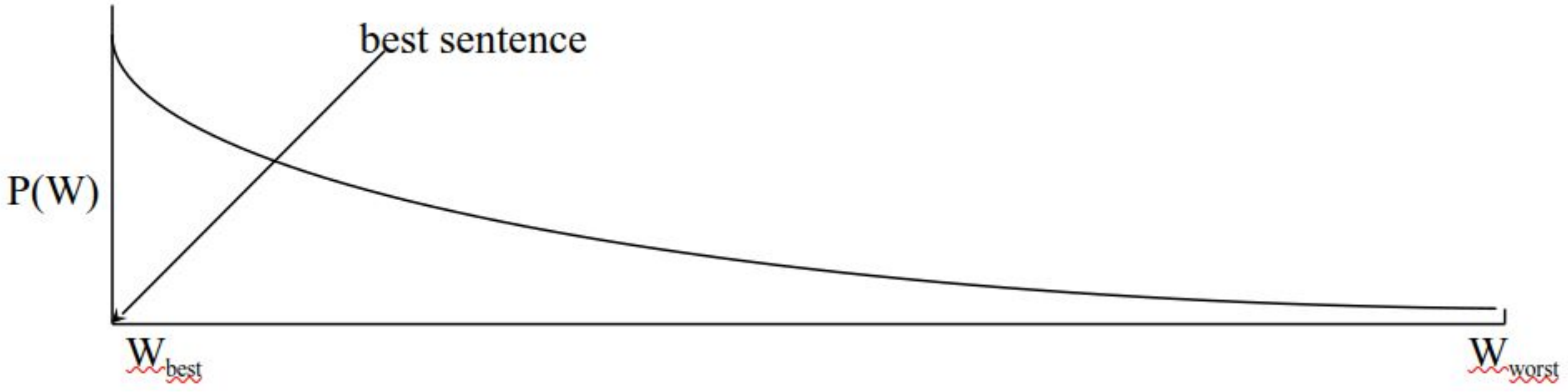
Imagine:

- Each sentence $W = \{ w_1, w_2, ..., w_n \}$ gets a probability $P(W|X)$ in a context X (think of it in the intuitive sense for now)
- For every possible context X, sort all the imaginable sentences W according to $P(W|X)$:
- Ideal situation:



Real World Situation

- Unable to specify set of grammatical sentences today using fixed “categorical” rules (maybe never, cf. arguments in M&S)
- Use statistical “model” based on **real world data** and care about the best sentence only (disregarding the “grammaticality” issue)



Probability

Experiments & (Finite) Sample Spaces

Sample space Ω – set of possible basic **outcomes**

- coin toss
- two 6-sided dice roll
- binary opinion poll, quality test
- lottery

- spelling errors
- next word

Experiments & (Finite) Sample Spaces

Sample space Ω – set of possible basic **outcomes**

- coin toss — Ω = {H, T}
- two 6-sided dice roll — Ω = {2 … 12}
- binary opinion poll, quality test — Ω = {yes, no}, Ω = {good, bad}
- lottery — Ω = {*an awful lot of stuff*}

- spelling errors — Ω = Z* *where Z is alphabet*
- next word — Ω = V *(supported vocabulary)*

Events

Event A – set of basic outcomes ($A \subseteq \Omega$)

- Certain event
- Impossible event

Events

Event A – set of basic outcomes ($A \subseteq \Omega$)

- Certain event, $A = \Omega$
- Impossible event, $A = \emptyset$

Example: 3x coin toss

- $\Omega = ?$

Event A – set of basic outcomes ($A \subseteq \Omega$)

- Certain event, $A = \Omega$
- Impossible event, $A = \varnothing$

Example: 3x coin toss

- Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}, $|\Omega| = 8$

Events

Event A – set of basic outcomes ($A \subseteq \Omega$)

- Certain event, $A = \Omega$
- Impossible event, $A = \varnothing$

Example: 3x coin toss

- Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}, $|\Omega| = 8$
- Possible events: “two heads”, “all tails”
    - A = ?
    - A = ?

Events

Event A – set of basic outcomes ($A \subseteq \Omega$)

- Certain event, $A = \Omega$
- Impossible event, $A = \varnothing$

Example: 3x coin toss

- Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}, $|\Omega| = 8$
- Possible events: “two heads”, “all tails”
    - A = {HHT, HTH, THH}
    - A = {TTT} (elementary event)

… what is the probability that these events happen?

Probability

Repeat experiment, **count** occurrences of event A

- Repeat in series, note the final count
- Divide by number of trials per series

- Result close to some unknown but **constant** value
- Call this **probability** of A, denote **p(A)**

Estimating Probability

- True probability **unknown**
- We can estimate from our observations
  - Either from a single series,
  - .. or take (weighted) average from each series
  - .. or concatenate all series

Estimating Probability

- True probability **unknown**
- We can estimate from our observations
  - Either from a single series,
  - .. or take (weighted) average from each series
  - .. or concatenate all series

$$p(A) = \frac{1}{N}\sum_{i=1}^{N} c_i/T_i = \frac{\sum_{i=1}^{N} c_i}{\sum_{i=1}^{N} T_i}$$

- Maximum Likelihood Estimate

*... this is the **best** estimate*

Probability Estimation

Example

- 3x coin toss
  - Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}, $|\Omega| = 8$
  - Count occurrences of the “two heads” event
  - A = {HHT, HTH, THH}

- Experiment outcomes
  - First series: run 1000 times, get 386 occurrences of A
  - Estimate p(A) = 0.386

  - Subsequent series results (all 1000 trials): 373, 399, 382, 355, 372, 406, 359
  - Estimate p(A) = 0.379

  - Assuming **uniform** distribution, p(A) = 3 / 8 = 0.375

Properties of Probability

Basic properties (also formal definition):

1. $0 \leq p(A) \leq 1$ *probability is between 0 and 1*
2. $p(\Omega) = 1$ *probability of certain event is 1*
3. $p(\cup(A_i)) = \sum p(A_i)$ *(only for disjoint events!)*

Consequences

- $p(\varnothing) = 0$
- $p(\bar{A}) = 1 - p(A)$
- $A \subseteq B \rightarrow p(A) \leq p(B)$ *(what about proper subset?)*
- $\sum_a p(a) = 1$

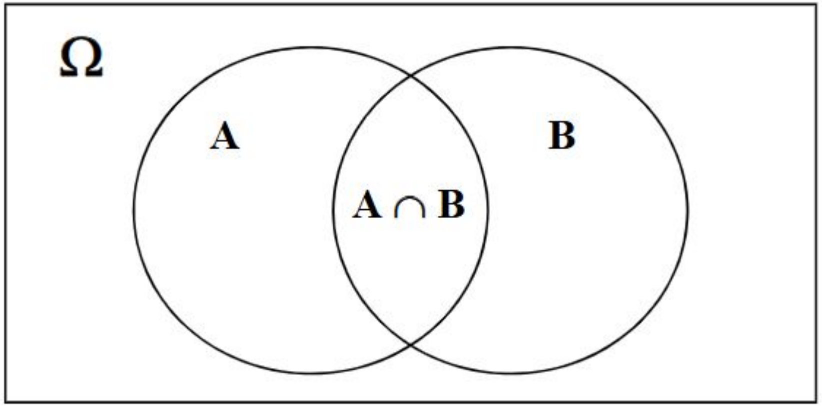
Joint and Conditional Probability

Combining probabilities of **multiple** events

- **Joint** probability: $p(A, B) = p(A \cap B)$
- **Conditional** probability: $p(A|B) = p(A, B) / p(B)$

When estimating from counts,

$$\begin{aligned} p(A|B) &= p(A, B) / p(B) \\ &= (c(A \cap B) / T) / (c(B) / T) \\ &= c(A \cap B) / c(B) \end{aligned}$$



Note: A and B can even be from different Ωs!

Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

.. follows from symmetry of joint probability.

Independence

Computing joint probability from the marginal distributions

- Can we calculate p(A,B) from p(A) and p(B)?

Independence

Computing joint probability from the marginal distributions

- Can we calculate p(A,B) from p(A) and p(B)?
- Using Bayes rule, we have

$$p(A|B) = p(B|A)\,p(A) / p(B)$$
$$p(A|B)\,p(B) = p(B|A)\,p(A)$$
$$p(A, B) = p(B|A)\,p(A)$$

Independence

Computing joint probability from the marginal distributions

- Can we calculate $p(A,B)$ from $p(A)$ and $p(B)$?
- Using Bayes rule, we have

$$p(A|B) = p(B|A)\, p(A) / p(B)$$
$$p(A|B)\, p(B) = p(B|A)\, p(A)$$
$$p(A, B) = p(B|A)\, p(A)$$

- How does $p(B|A)$ relate to $p(B)$?

Independence

Computing joint probability from the marginal distributions

- Can we calculate p(A,B) from p(A) and p(B)?
- Using Bayes rule, we have

$$p(A|B) = p(B|A)\,p(A) / p(B)$$
$$p(A|B)\,p(B) = p(B|A)\,p(A)$$
$$p(A, B) = p(B|A)\,p(A)$$

- How does p(B|A) relate to p(B)?
  - Does knowing A tell us something about B?
  - If **not**, p(B|A) = p(B), and we say that A and B are **independent.**
  - In this case, p(A, B) = p(A) * p(B)

Independence

Computing joint probability from the marginal distributions

- Can we calculate p(A,B) from p(A) and p(B)?
- Using Bayes rule, we have

$$p(A|B) = p(B|A)\,p(A) / p(B)$$
$$p(A|B)\,p(B) = p(B|A)\,p(A)$$
$$p(A, B) = p(B|A)\,p(A)$$

- How does p(B|A) relate to p(B)?
  - Does knowing A tell us something about B?
  - If **not**, p(B|A) = p(B), and we say that A and B are **independent.**
  - In this case, p(A, B) = p(A) * p(B)

Examples: two coin tosses, weather conditions years apart, ...

Chain Rule

Joint and conditional probabilities for **many events**

$$p(A_1, A_2, \ldots, A_n) = p(A_1|A_2, A_3, \ldots, A_n) \times$$
$$\times p(A_2|A_3, \ldots, A_n) \times \ldots \times p(A_{n-1}|A_n) \times p(A_n)$$

.. useful in NLP where we can approximate some of the terms

The “Golden Rule” of Classic Statistical NLP

P(A|B) in NLP applications

- Speech recognition, machine translation, language modeling
  - B = audio signal, source sentence, previous word
  - A = transcription, target sentence, next word

The "Golden Rule" of Classic Statistical NLP

P(A|B) in NLP applications

- Speech recognition, machine translation, language modeling
  - B = audio signal, source sentence, previous word
  - A = transcription, target sentence, next word
- Goal is to find A that **maximizes P(A|B)**

$$A^* = \text{argmax}_A \, p(A|B)$$

The “Golden Rule” of Classic Statistical NLP

P(A|B) in NLP applications

- Speech recognition, machine translation, language modeling
  - B = audio signal, source sentence, previous word
  - A = transcription, target sentence, next word
- Goal is to find A that **maximizes P(A|B)**

$$A^* = \text{argmax}_A\, p(A|B)$$

- When estimating P(A|B) directly is not desirable, use Bayes rule

$$A^* = \text{argmax}_A\, p(B|A)p(A)\cancel{/p(B)}$$

- Ignore the p(B) term which is **constant**

Random Variables

Statistical outcomes with **numeric values**

- **X** is a function from Ω, returns a value, typically a real number
- Simplify real world into a model (throwing wooden dice → numbers)
- Can also return items from finite set → *discrete R.V.*

Random Variables

Statistical outcomes with **numeric values**

- **X** is a function from Ω, returns a value, typically a real number
- Simplify real world into a model (throwing wooden dice → numbers)
- Can also return items from finite set → *discrete R.V.*

Examples

- 6-sided die → numbers from 1 to 6
- Coin toss → 0 or 1

Random Variables

Statistical outcomes with **numeric values**

- **X** is a function from Ω, returns a value, typically a real number
- Simplify real world into a model (throwing wooden dice → numbers)
- Can also return items from finite set → *discrete R.V.*

Examples

- 6-sided die → numbers from 1 to 6
- Coin toss → 0 or 1

Notation: $p_X(x)$, $p(X=x)$, or just $p(x)$ if the context is clear.

Joint and Conditional Distributions for RVs

Properties of joint and conditional RVs similar to events

- $p_{X|Y}(x, y) = p_{XY}(x|y) = p(x|y)$ (various notation)
- $p(x|y) = p(y|x)\, p(x) / p(y)$ Bayes rule
- $p(x,y,z) = p(x|y,z)\, p(y|z)\, p(z)$ Chain rule

Expectation

Mean value of a random variable

- Average of all possible values, weighted by their probabilities

$$E(X) = \sum_{x \in X(\Omega)} x \cdot p(X = x)$$

Examples

- 6-sided die → 3.5
- two dice → 7
- coin toss → 0.5

Binomial Distribution

- trial outcome: 0 or 1 (thus: **bi**nomial)
- make n trials
- interested in the number of successes r or probability of a success p
- Formally: $B(n,p)$

Moodle Quiz

Moodle Quiz

https://dl1.cuni.cz/course/view.php?id=18547