Pattern Analysis In Natural Language Texts

English for Computer Science 1

TECHNISCHE UNIVERSITÄT DARMSTADT

Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

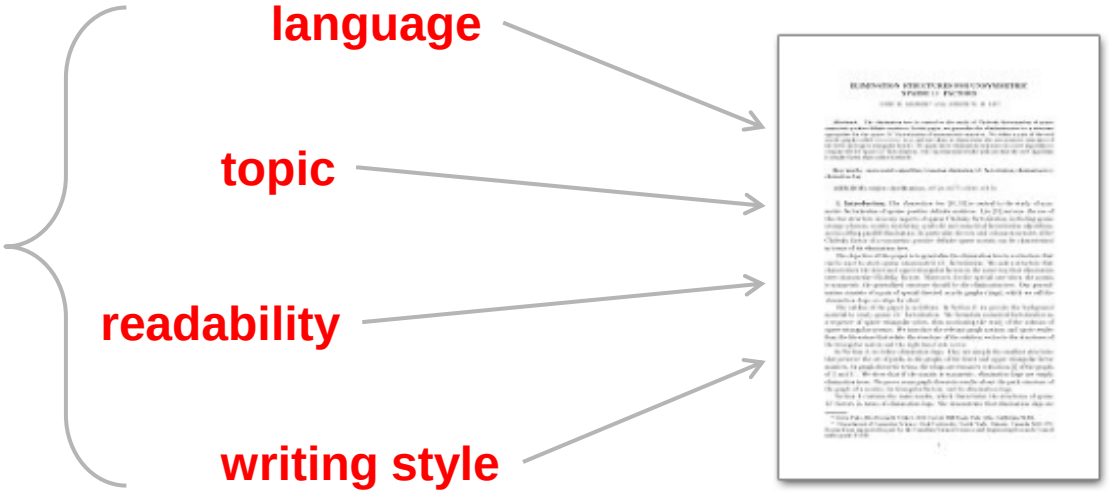
- Motivation
- Applications
- Introduction to Pattern Analysis
- Take-Home-Messages
- References

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Natural language texts often provide valuable information (beyond the semantics of the text)
- What does it mean?
- Texts typically consists of many “surface properties”, e.g.

Such properties are usefull for further investigation…



Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- How difficult is the task of retrieving such properties?
  - For humans: easy (just "read" the text…)
  - For machines: even more easier… (why?)

As far as we know: "Machines don't understand language"

However, they don't have to in order to solve the task…

→ focus of this presentation !

Applications

TECHNISCHE UNIVERSITÄT DARMSTADT

- Author identification (e.g. forensic related documents)
- Measuring level of education (readability tests…)
- Cryptanalysis (for simple substitution ciphers)
- Reconstruct fragmentary texts (known as „Noisy Channel Model“)
- Finding similarities between texts (duplicate detection / plagiarism)
- …and many more ;-)

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- Differences between humans and machines for the task

→ Humans: slow but higher precision

→ Machines: fast but error-prone

- Who should solve the task?

  Depends on the task…

  - Analyzing few (critical) texts should be judged by humans
  - Analyzing billions of (harmless) texts should be done automaticaly

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Sample text:

After being released from prison in 2009, Levi fulfilled a promise to himself to visit Israel. He arrived in September, underwent a formal conversion and began to study Judaism with rabbis associated with the more stringent ultra-orthodox sects of the faith. He says he wanted to learn more about the religion and the meaning behind its many rituals…

What kind of properties can you infer from this text?

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Sample text:

After being released from prison in 2009, Levi fulfilled a promise to himself to visit Israel. He arrived in September, underwent a formal conversion and began to study Judaism with rabbis associated with the more stringent ultra-orthodox sects of the faith. He says he wanted to learn more about the religion and the meaning behind its many rituals…

- We can observe several properties, e.g.

  Language: english

  Topic: religion

  Type: news text (what is the indicator?)

  Readability: easy, but requires background knowledge (rabbis, Judaism, …)

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Example (I) Identify language of a given text

Statistical approach: determine the most frequent n-gram‘s in the text…

Sample word: “cats”

Name	**n**	**n-gram‘s**
Unigram	n = 1	{ c, a, t, s }
Bigram	n = 2	{ ca, at, ts }
Trigram	n = 3	{ cat, ats }
Tetragram	n = 4	{ cats }

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Consider again the sample text:

After being released from prison in 2009, Levi fulfilled a promise to himself to visit Israel. He arrived in September, underwent a formal conversion and began to study Judaism with rabbis associated with the more stringent ultra-orthodox sects of the faith. He says he wanted to learn more about the religion and the meaning behind its many rituals…

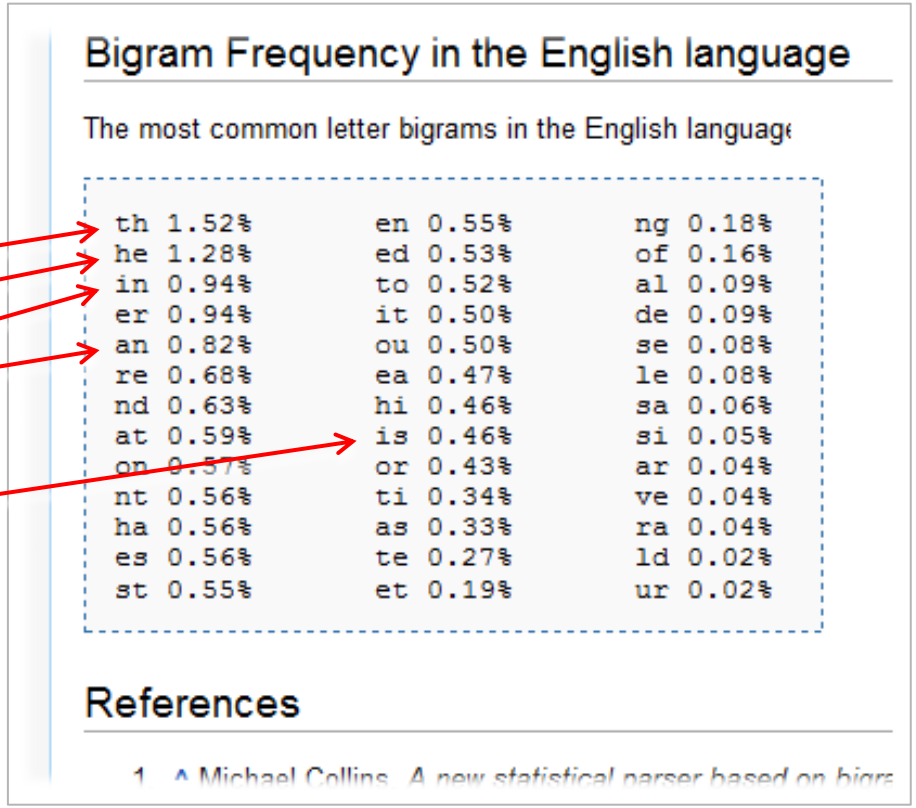
The 5 frequent bigrams here are: { th, he, an, in, is }

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

The 5 frequent bigrams:

{
th,
he,
an,
in,
is
}



Bigram Frequency in the English language

The most common letter bigrams in the English languag

th 1.52%	en 0.55%	ng 0.18%
he 1.28%	ed 0.53%	of 0.16%
in 0.94%	to 0.52%	al 0.09%
er 0.94%	it 0.50%	de 0.09%
an 0.82%	ou 0.50%	se 0.08%
re 0.68%	ea 0.47%	le 0.08%
nd 0.63%	hi 0.46%	sa 0.06%
at 0.59%	is 0.46%	si 0.05%
on 0.57%	or 0.43%	ar 0.04%
nt 0.56%	ti 0.34%	ve 0.04%
ha 0.56%	as 0.33%	ra 0.04%
es 0.56%	te 0.27%	ld 0.02%
st 0.55%	et 0.19%	ur 0.02%

References

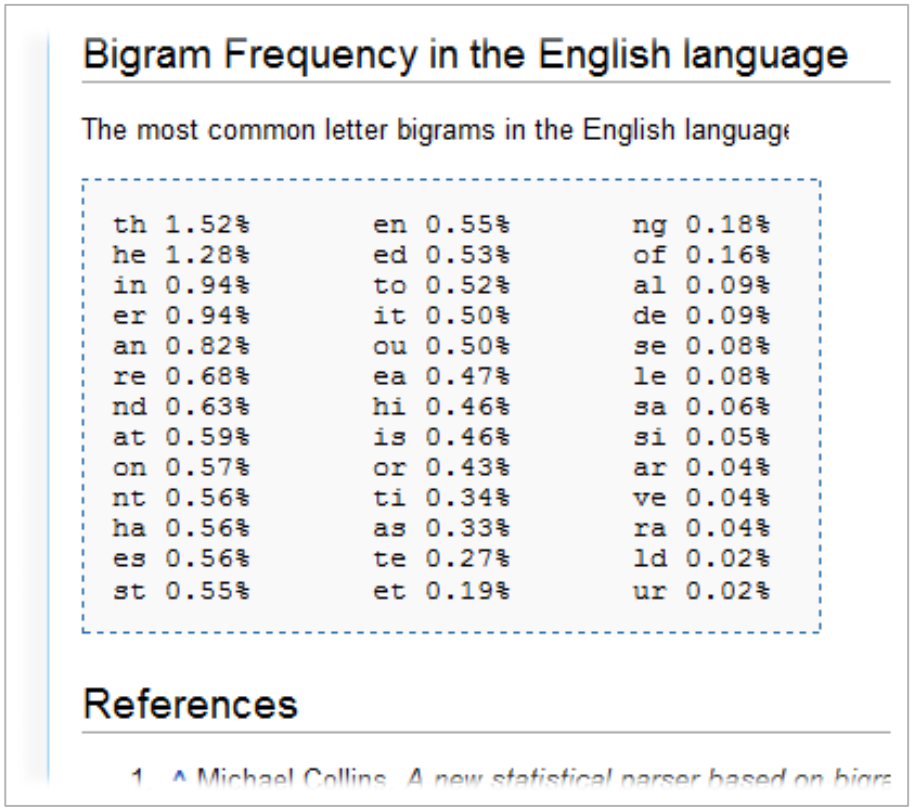
1. ^ Michael Collins. *A new statistical parser based on bigra*

Source: [http://en.wikipedia.org/wiki/Bigram]

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Why is the text assumed to be written in English?



Bigram Frequency in the English language

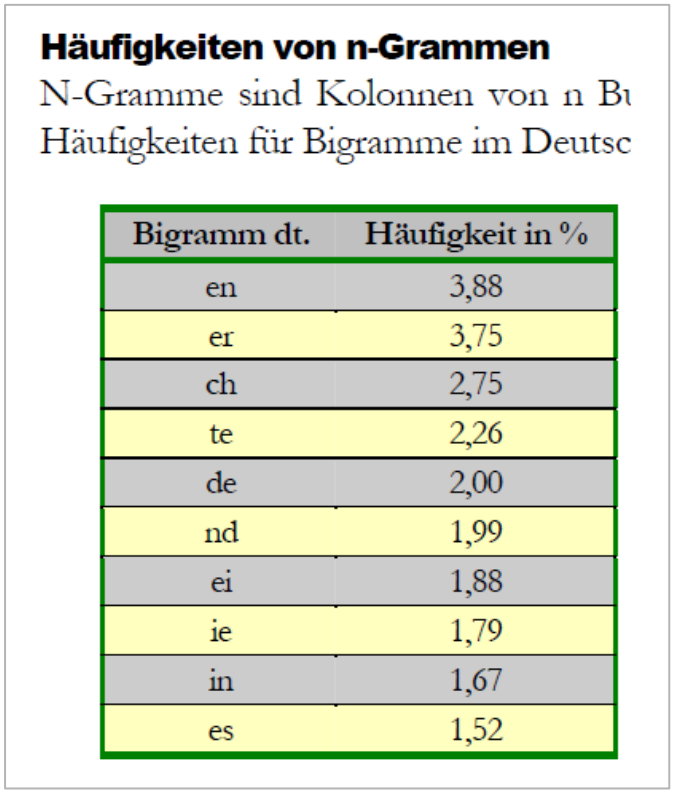
The most common letter bigrams in the English language

th 1.52%	en 0.55%	ng 0.18%
he 1.28%	ed 0.53%	of 0.16%
in 0.94%	to 0.52%	al 0.09%
er 0.94%	it 0.50%	de 0.09%
an 0.82%	ou 0.50%	se 0.08%
re 0.68%	ea 0.47%	le 0.08%
nd 0.63%	hi 0.46%	sa 0.06%
at 0.59%	is 0.46%	si 0.05%
on 0.57%	or 0.43%	ar 0.04%
nt 0.56%	ti 0.34%	ve 0.04%
ha 0.56%	as 0.33%	ra 0.04%
es 0.56%	te 0.27%	ld 0.02%
st 0.55%	et 0.19%	ur 0.02%

References

1. ^ Michael Collins. *A new statistical parser based on bigra*

Source: [http://en.wikipedia.org/wiki/Bigram]



Häufigkeiten von n-Grammen

N-Gramme sind Kolonnen von n Bu
Häufigkeiten für Bigramme im Deutsc

Bigramm dt.	Häufigkeit in %
en	3,88
er	3,75
ch	2,75
te	2,26
de	2,00
nd	1,99
ei	1,88
ie	1,79
in	1,67
es	1,52

Source: [http://tinyurl.com/296z5wr]

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Example (II) Predict the topic for a given text

- Statistical model: Naïve Bayes classification
- Naïve assumption:

> "All words in a given text are conditionally independent from eachother"

- **Note:** assumption rarely holds in natural language texts…
- Goal:

> Determine the probability p(c|d) that document d belongs to category c

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- How the model works…
  - Precondition: trainingset **D** (labeled sample texts)
  - Anatomy of the model:

Category (e.g. sports, politics, economy,…)

$$c = \arg\max_c \; p(c) \prod_{i=1}^{|\mathbf{d}|} p(t_i|c)$$

Number of all words (t_i's) in the unknown text **d**

Fraction of documents that belongs to category c

Probability that word t_i occurs at a certain position in the document

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Trainingset **D** = {d_1 , d_2, d_3, d_4}

Document d_i	**Contents**	**Category**
d_1	chinese beijing chinese	C
d_2	chinese chinese shanghai	C
d_3	chinese macao	C
d_4	tokyo yapan chinese	J

Vocabulary **V** = { beijing, chinese, macao, shanghai, tokyo, yapan }

New (unknown) document d_5 = "chinese, chinese, chinese, tokyo, yapan"

Predict which category d_5 most likely belongs to…

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

New (unknown) document d_5 = "chinese, chinese, chinese, tokyo, yapan"

Model needs to compute: P(C), P(J) and $P(t_i|C)$, $P(t_i|J)$

$$P(C) = \frac{\text{amount of documents of category } C \text{ in } D}{\text{amount of documents in } D} = \frac{3}{4}$$

$$P(J) = \frac{\text{amount of documents of category } J \text{ in } D}{\text{amount of documents in } D} = \frac{1}{4}$$

$$P(chinese \mid C) = \frac{\text{occurrence of "chinese" in } C + 1}{sum(\text{occurrence of each word in } C) + |V|} = \frac{5+1}{8+6} = \frac{3}{7}$$

Introduction to Pattern Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

All the parameters:

P(chinese|C) = 3/7

P(tokyo|C) = 1/14

P(yapan|C) = 1/14

P(chinese|J) = 2/9

P(tokyo|J) = 2/9

P(yapan|J) = 2/9

$$P(d_5|C) = P(C)\,P(chinese|C)^3 P(tokyo|C) P(yapan|C)$$

$$P(d_5|J) = P(J)\,P(chinese|J)^3 P(tokyo|J) P(yapan|J)$$

$$P(d_5|C) = 0.000301...$$

$$P(d_5|J) = 0.000135...$$

Hence, category C would best fit for d_5

Or with other words: d_5 is labeled with the topic C

Take-Home-Messages

TECHNISCHE UNIVERSITÄT DARMSTADT

Pattern Analysis in natural language texts:

- …is very valuable (if you know where and how to use it)
- …becomes very popular in many fields (beyond computer science)
- …doesn't require semantics of text (hence, easy to implement)
- …is not 100% bulletproof, but often the last possibility to gain insights about the text

- Last but not least: it's a  (hot) discussed research topic !!!

TECHNISCHE UNIVERSITÄT DARMSTADT

Your turn ;-)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Imagine you want to detect plagiarism between two documents (A and B)

  Explain: which of the following 2 ideas would you choose for the detection?

  (1) Overlap in sets of words

  (2) Overlap in sequences of words (n-grams)

- Recall the calculation:

$$P(chinese \mid C) = \frac{\text{occurrence of "chinese" in } C + 1}{sum(\text{occurrence of each word in } C) + |V|}$$

  What can we observe, if we ignore the red part in the fraction?

- The presentation mentioned the "Readability" property, how could a machine compute this?

TECHNISCHE UNIVERSITÄT DARMSTADT

Thanks for your attention…

References

- "**Foundations of Statistical Natural Language Processing**",

  http://nlp.stanford.edu/fsnlp

- "**Einführung in die Kryptologie**",

  Tino Hempel,
  Einführung in die Kryptologie, (Fach Mathematik), 1995.

  http://www.tinohempel.de/info/info/kryptografie/download/krypto.pdf

- "**Questions picture**",

  Photo of "Le Penseur", a bronze sculpture made by Auguste Rodin,
  held in the Musée Rodin in Paris, France.

- "**Rap bad boy goes kosher**",

  CNN article (November 18, 2010|By Izzy Lemberg and Kevin Flower),

  http://bit.ly/koW3MM