THE UNIVERSITY of ADELAIDE

CRICOS PROVIDER 00123M

ANLP week1: Introduction part 1

Dr Alfred Krzywicki
University of Adelaide

Welcome to Week 1 of Applied NLP!

Natural Language Processing

"*Natural language processing is a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like languages processing for a range of particular tasks or applications*."
by Liddy (1998)

Some other names:

Computational Linguistics
Natural Language Engineering
Speech and Text Processing

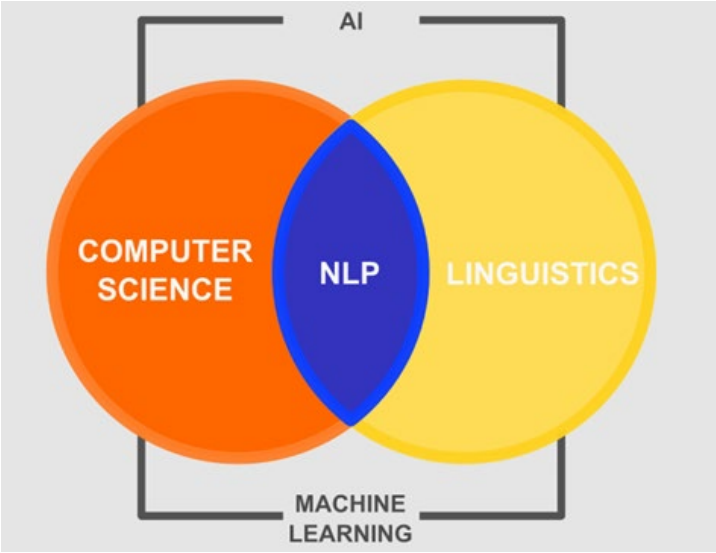
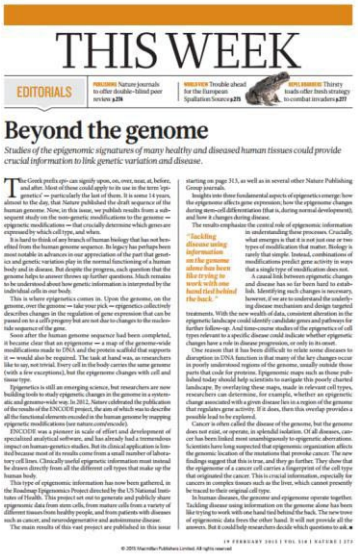


Image source: algorithmxlab.com

Language

A vocabulary consists of a set of words (w_i)

A text is composed of a sequence of words from a vocabulary



THIS WEEK

EDITORIALS

PUBLISHING Nature journals to offer double-blind peer review p.278

WILDLIFE Trouble ahead for the European Spallation Source p.271

INVASIVE SPECIES Thirty leads offer Irish strategy to combat invaders p.277

Beyond the genome

Studies of the epigenomic signatures of many healthy and diseased human tissues could provide crucial information to link genetic variation and disease.

The Greek prefix *epi-* can signify upon, on, over, near, at, before, and after. Most of those could apply to its use in the term 'epigenetics' — particularly the last of them. It is some 14 years, almost to the day, that Nature published the draft sequence of the human genome. Now, in this issue, we publish results from a subsequent study on the non-genetic modifications to the genome — epigenetic modifications — that crucially determine which genes are expressed by which cell type, and when.

It is hard to think of any branch of human biology that has not benefited from the human genome sequence. Its legacy has perhaps been most notable in advances in our appreciation of the part that genetics and genetic variation play in the normal functioning of a human body and in disease. But despite the progress, each question that the genome helps to answer throws up further questions. Much remains to be understood about how genetic information is interpreted by the individual cells in our body.

This is where epigenetics comes in. Upon the genome, on the genome, over the genome — take your pick — epigenetics collectively describes changes in the regulation of gene expression that can be passed on to a cell's progeny but are not due to changes to the nucleotide sequence of the gene.

Soon after the human genome sequence had been completed, it became clear that an epigenome — a map of the genome-wide modifications made to DNA and the protein scaffold that supports it — would also be required. The task at hand was, as researchers like to say, not trivial. Every cell in the body carries the same genome (with a few exceptions), but the epigenome changes with cell and tissue type.

Epigenetics is still an emerging science, but researchers are now building tools to study epigenetic changes in the genome in a systematic and genome-wide way. In 2012, Nature celebrated the publication of the results of the ENCODE project, the aim of which was to describe all the functional elements encoded in the human genome by mapping epigenetic modifications (see nature.com/encode).

ENCODE was a pioneer in scale of effort and development of specialized analytical software, and has already had a tremendous impact on human-genetics studies. But its clinical application is limited because most of its results come from a small number of laboratory cell lines. Clinically useful epigenetic information must instead be drawn directly from all the different cell types that make up the human body.

This type of epigenomic information has now been gathered, in the Roadmap Epigenomics Project directed by the US National Institutes of Health. This project set out to generate and publicly share epigenomic data from stem cells, from mature cells from a variety of different tissues from healthy people, and from patients with diseases such as cancer, and neurodegenerative and autoimmune disease.

The main results of this vast project are published in this issue starting on page 313, as well as in several other Nature Publishing Group journals.

Insights into three fundamental aspects of epigenetics emerge: how the epigenome affects gene expression, how the epigenome changes during stem-cell differentiation (that is, during normal development), and how it changes during disease.

The results emphasize the central role of epigenomic information in understanding these processes. Crucially, what emerges is that it is not just one or two types of modification that matter. Biology is rarely that simple. Instead, combinations of modifications predict gene activity in ways that a single type of modification does not.

A causal link between epigenetic changes and disease has so far been hard to establish. Identifying such changes is necessary, however, if we are to understand the underlying disease mechanism and design targeted treatments. With the new wealth of data, consistent alteration in the epigenetic landscape could identify candidate genes and pathways for further follow-up. And time-course studies of the epigenetics of cell types relevant to a specific disease could indicate whether epigenetic changes have a role in disease progression, or only in its onset.

> "Tackling disease using information on the genome alone has been like trying to work with one hand tied behind the back."

One reason that it has been difficult to relate some diseases to disruption in DNA function is that many of the key changes occur in poorly understood regions of the genome, usually outside those parts that code for proteins. Epigenomic maps such as those published today should help scientists to navigate this poorly charted landscape. By overlaying these maps, made in relevant cell types, researchers can determine, for example, whether an epigenetic change associated with a given disease lies in a region of the genome that regulates gene activity. If it does, then this overlap provides a possible lead to be explored.

Cancer is often called the disease of the genome, but the genome does not exist, or operate, in splendid isolation. Of all diseases, cancer has been linked most unambiguously to epigenetic aberrations. Scientists have long suspected that epigenomic organization affects the genomic location of the mutations that provoke cancer. The new findings suggest that this is true, and they go further. They show that the epigenome of a cancer cell carries a fingerprint of the cell type that originated the cancer. This is crucial information, especially for cancers in complex tissues such as the liver, which cannot presently be traced to their original cell type.

In human diseases, the genome and epigenome operate together. Tackling disease using information on the genome alone has been like trying to work with one hand tied behind the back. The new trove of epigenomic data frees the other hand. It will not provide all the answers. But it could help researchers decide which questions to ask. ■

19 FEBRUARY 2015 | VOL 518 | NATURE | 273

© 2015 Macmillan Publishers Limited. All rights reserved

A language is constructed of a set of all possible texts

(http://www.old-engli.sh/language.php)

Artificial Language

```
try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.g
            acWriter.println(cMessage);
            acWriter.flush();
        } catch(IO
            System
            queue.
        }
    }
} catch(Interrupte
```

```
append(CharSequenc
append(char c)
append(CharSequenc
format(String form
format(Locale l, S
printf(String form
```

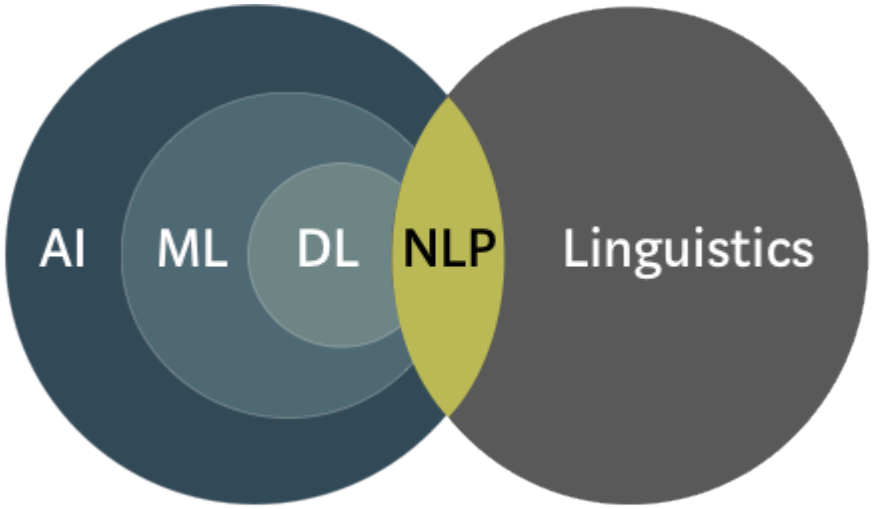
(https://netbeans.org/features/java/)

```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '    %s [label="%s' % (nodename,label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s"];' % ast[1]
        else:
            print '"]'
    else:
        print '"];'
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print '    %s -> {' % nodename,
        for name in children:
            print '%s' % name,
```

(http://noobite.com/learn-programming-start-with-python/)

Natural Language Processing



Natural language processing (NLP): how to program computers to process and analyse large amounts of natural language data.

Why NLP is hard?

- Natural language is highly ambiguous

"At last, a computer that understands you like your mother"

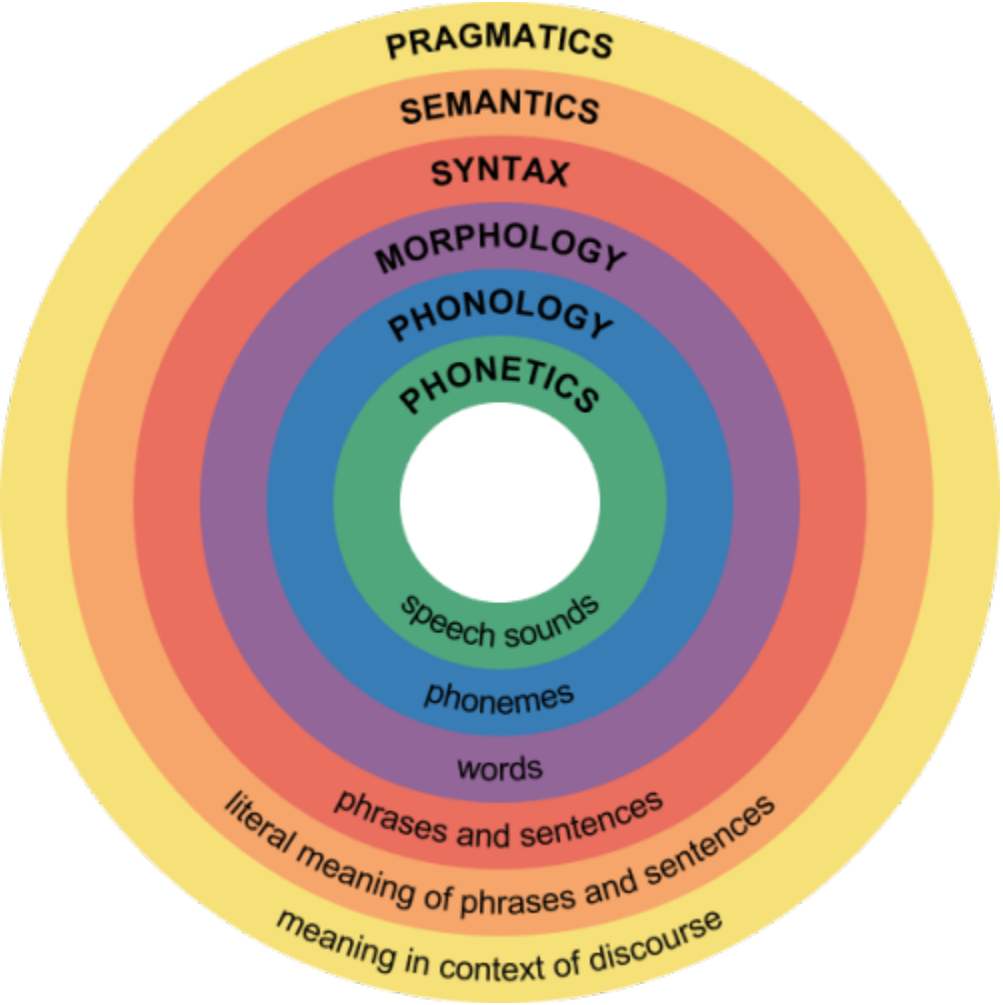
1. It understands you as well as your mother understands you
2. It understands (that) you like your mother
3. It understands you as well as it understands your mother

How many meanings can you recognize in this sentence?

One morning I shot an elephant in my pajamas.

Linguistic components

Sound ->
phoneme ->
morpheme ->
sub-word ->
word ->
phrase ->
sentence ->
paragraph ->
text ->
corpus



https://pediaa.com/difference-between-semantics-and-pragmatics/

Definitions of some Linguistic terms

- Phoneme a unit of sound that can distinguish one word from another in a particular language, e.g. the word 'cat' has three phonemes: /c/ /a/ /t/.
- Morpheme the smallest meaningful lexical item in a language. For example, "un-", "break", and "-able" in the word "unbreakable
- Sub-word just a part of word, like morphemes.
- Phrase a group of words or singular word acting as a grammatical unit. For example, adjective phrase "very happy“.
- Corpus a language resource consisting of a large and structured set of texts, e.g., a set of Wikipedia articles.
- Semantics meaning in a literal sense, e.g., “5 pm” means literally a time point.
- Pragmatics is the meaning in context, e.g., “5 pm” may mean it is time to go home.

https://en.wikipedia.org/wiki/
http://www.cse.unsw.edu.au/~billw/nlpdict.html

Parts of Speech

https://www.sketchengine.eu/penn-treebank-tagset/

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
IN	preposition, subordinating conjunction	in, of, like
JJ	adjective	green
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PP	personal pronoun	I, he, it
VB	verb be, base form	be

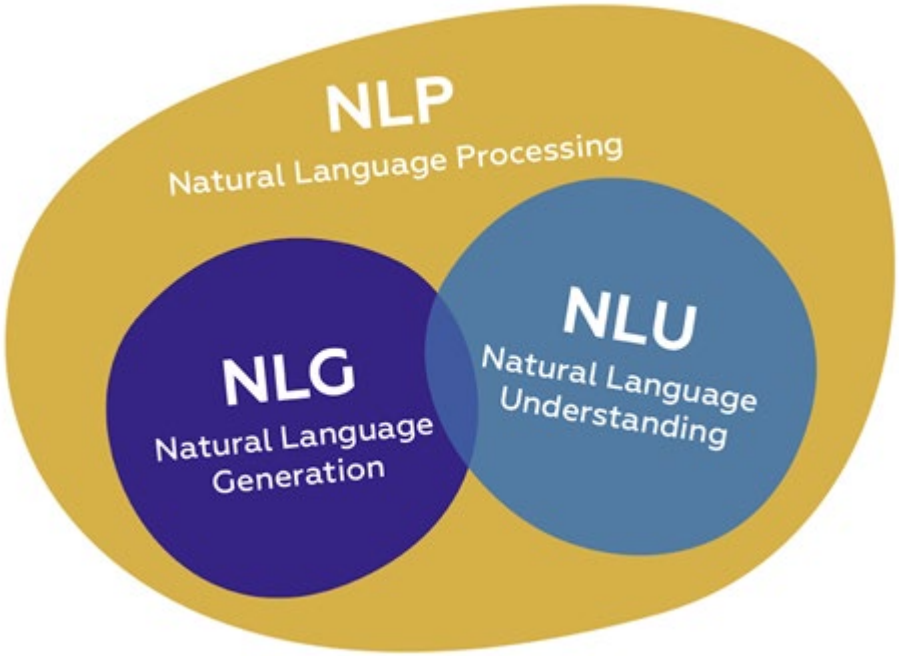
Natural Language Processing

NLU

- Syntactic parsing
- Coreference resolution (“Elon Musk”<=>”Tesla CEO)”
- Semantic parsing (meaning)
- Part-of-speech tagging (POS)
- Named entity recognition (NER) (“”Rome” ⇔ ”capital of Italy”
- Natural language inference (true, neutral, false)
- Relation extraction (“player wins game”)
- Text categorization
- Sentiment analysis
- ….

NLU & NLG

- Paraphrase
- Dialogue agents
- Question answering
- Text summarization
- Machine translation
- …



Text classification

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Nature. 2014 Mar 20;507(7492):323-8. doi: 10.1038/nature13145. Epub 2014 Mar 12.

Coupling of angiogenesis and osteogenesis by a specific vessel subtype in bone.

Kusumbe AP[1], Ramasamy SK[1], Adams RH[2].

Author information

Abstract

The mammalian skeletal system harbours a hierarchical system of mesenchymal stem cells, ost
bone formation. Osteogenesis is indispensable for the homeostatic renewal of bone as
frequently decline in ageing organisms, leading to loss of bone mass and increa
vessels in bone and osteogenesis are coupled, but relatively little is kno
a new capillary subtype in the murine skeletal system with distin
specific locations, mediate growth of the bone vasculatu
osteoprogenitors and couple angiogenesis to os
in bone from aged animals, and pharma

Comment in

Bone biology: Ve

PM

MeSH Terms

Aging/metabolism
Aging/pathology
Animals
Blood Vessels/
Blood

Sentiment Analysis

Customer Reviews

Speech and Language Processing, 2nd Edition

15 Reviews

5 star: (8)
4 star: (3)
3 star: (3)
2 star: (0)
1 star: (1)

Average Customer Review

★★★★☆ (15 customer reviews)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

4 of 4 people found the following review helpful

★★★★★ **Great introductions and reference book**

I read the first edition of that book and it is terrific. The second edition is much more adapted to current research. Statistical methods in NLP are more detailed and some syntax-based approaches are presented. My specific interest is in machine translation and dialogue systems. Both chapters are extensively rewritten and much more elaborated. I believe this book is...

Read the full review ›

Published on August 9, 2008 by carheg

› See more **5 star**, **4 star** reviews

Vs.

The most helpful critical review

37 of 37 people found the following review helpful

★★★☆☆ **Good description of the problems in the field, but look elsewhere for practical solutions**

The authors have the challenge of covering a vast area, and they do a good job of highlighting the hard problems within individual sub-fields, such as machine translation. The availability of an accompanying Web site is a strong plus, as is the extensive bibliography, which also includes links to freely available software and resources.

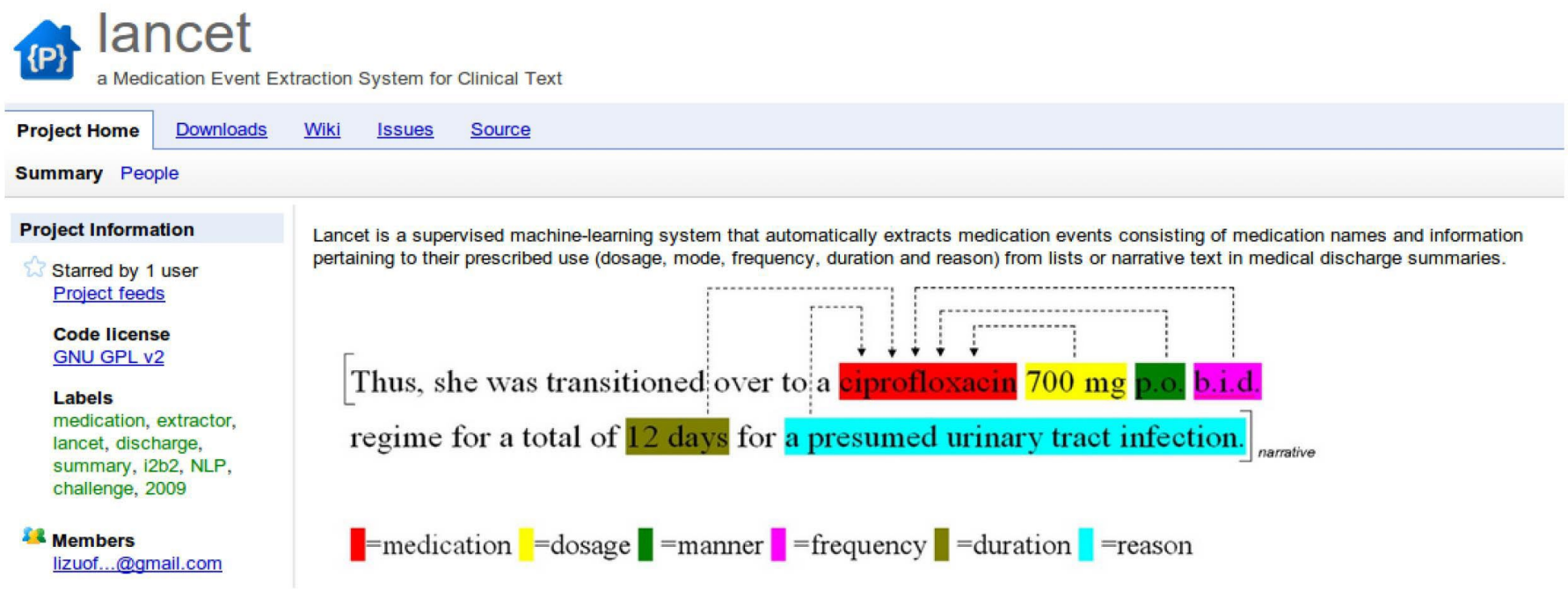
Now for the...

Read the full review ›

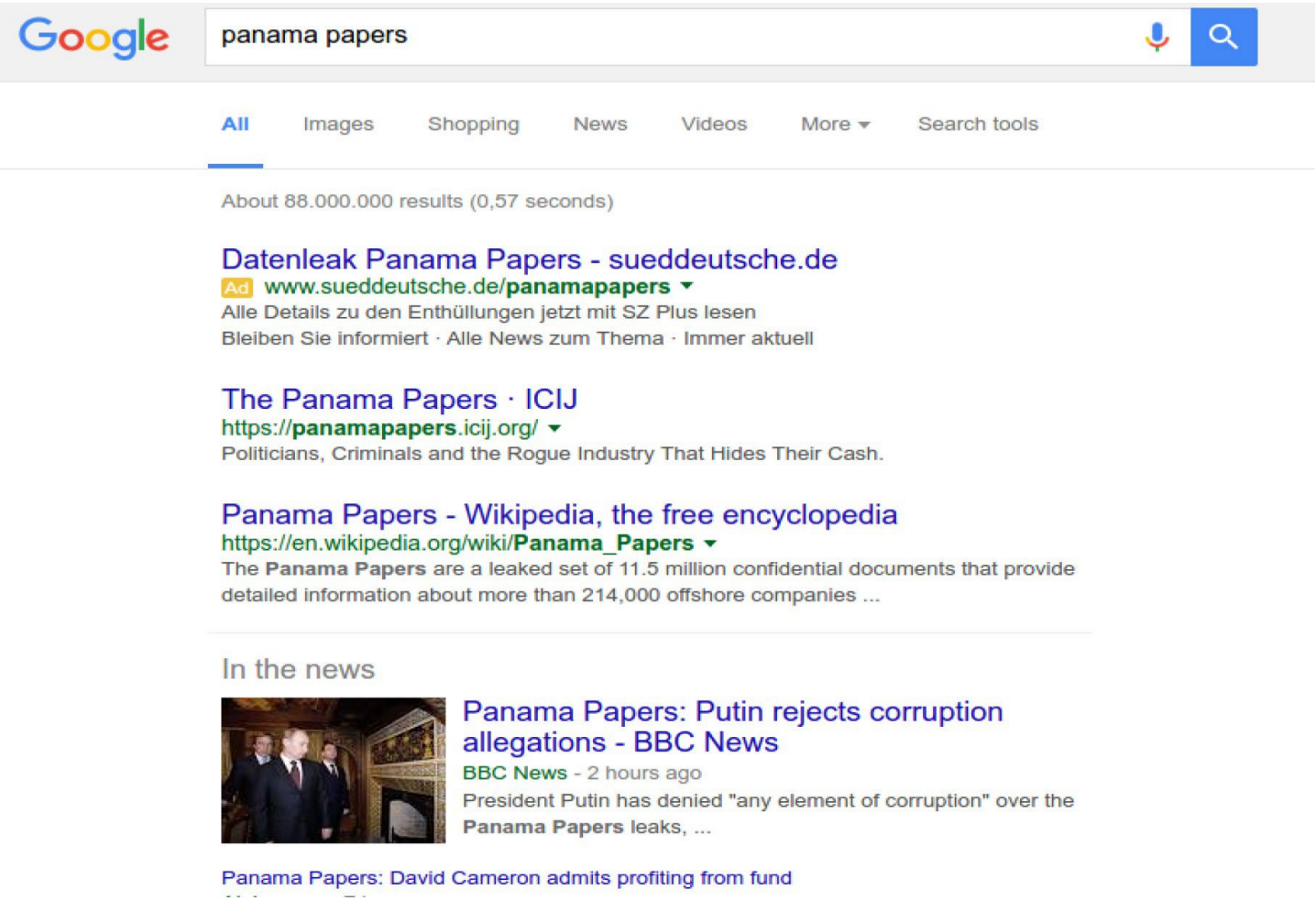
Published on April 2, 2009 by P. Nadkarni

› See more **3 star**, 2 star, **1 star** reviews

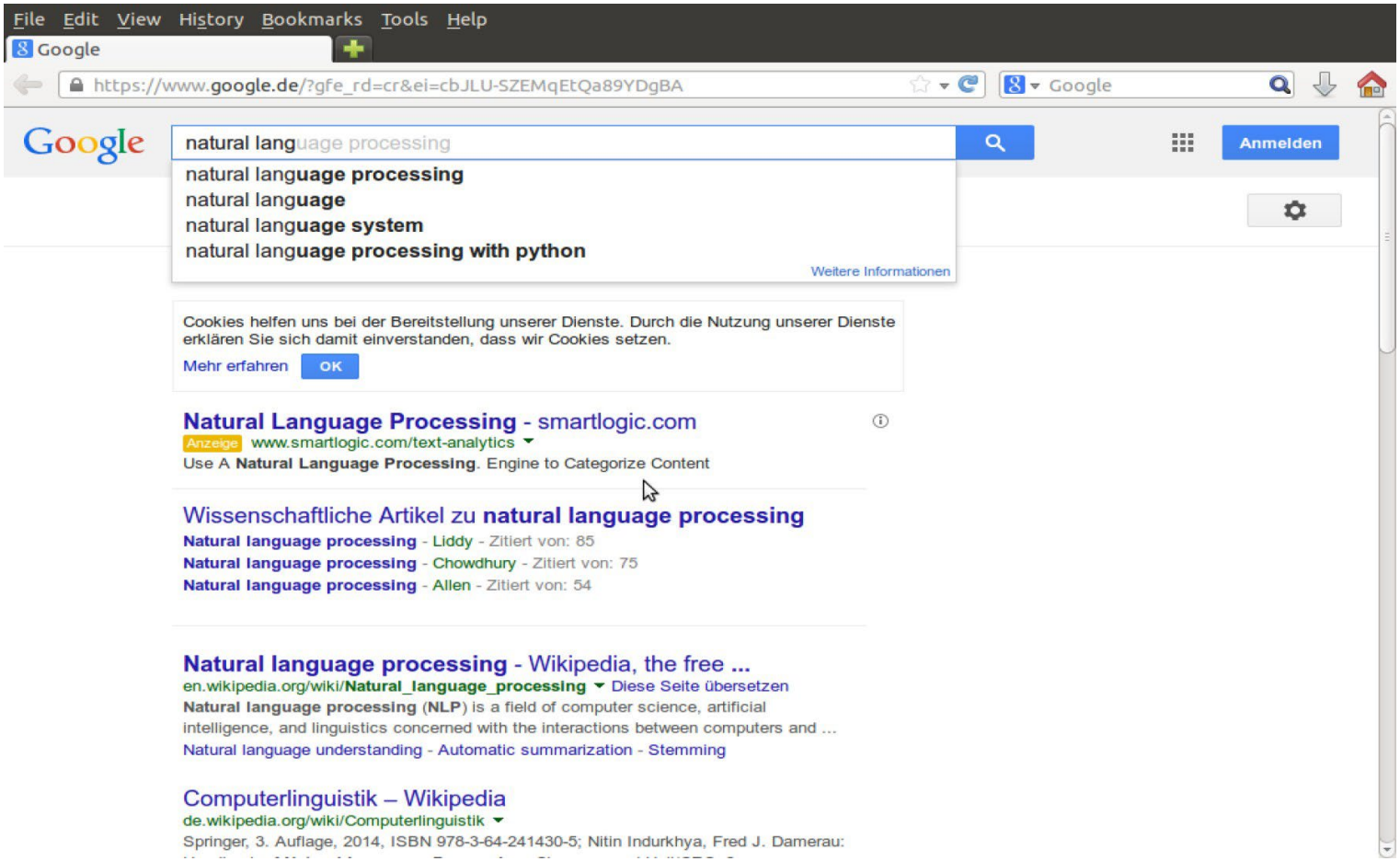
Information Extraction



Information retrieval

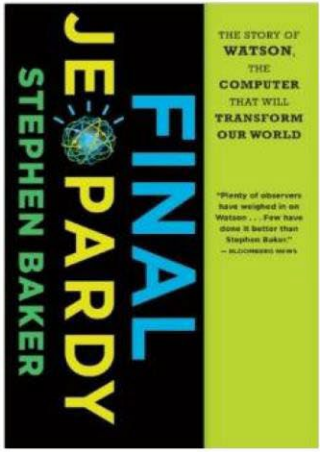


Word prediction



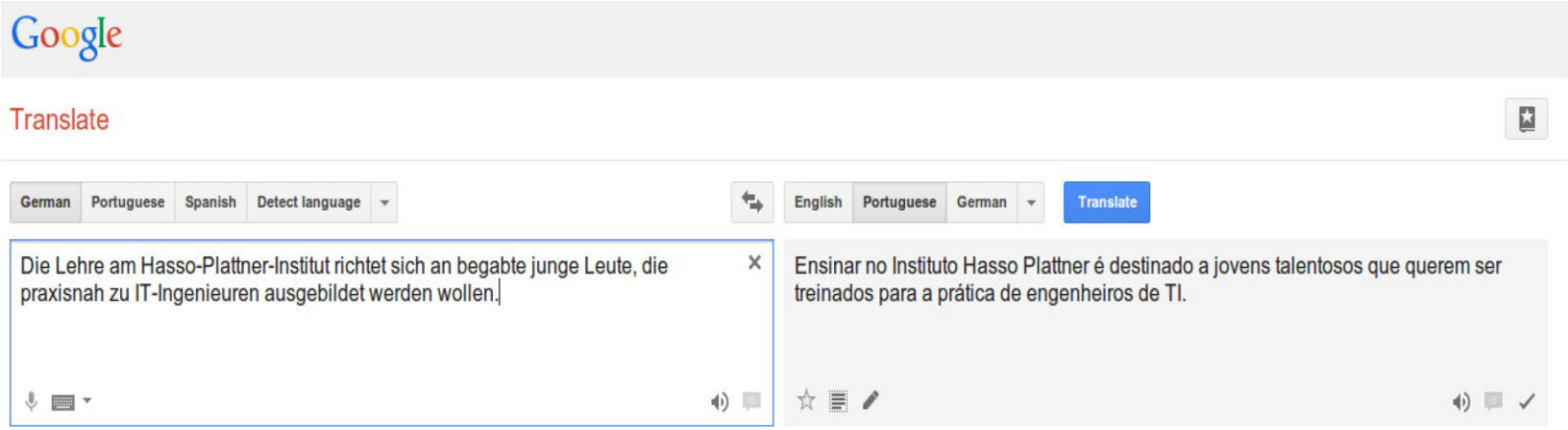
Question Answering

- IBM Watson in Jeopardy



https://www.youtube.com/watch?v=WFR3lOm_xhE

Machine Translation



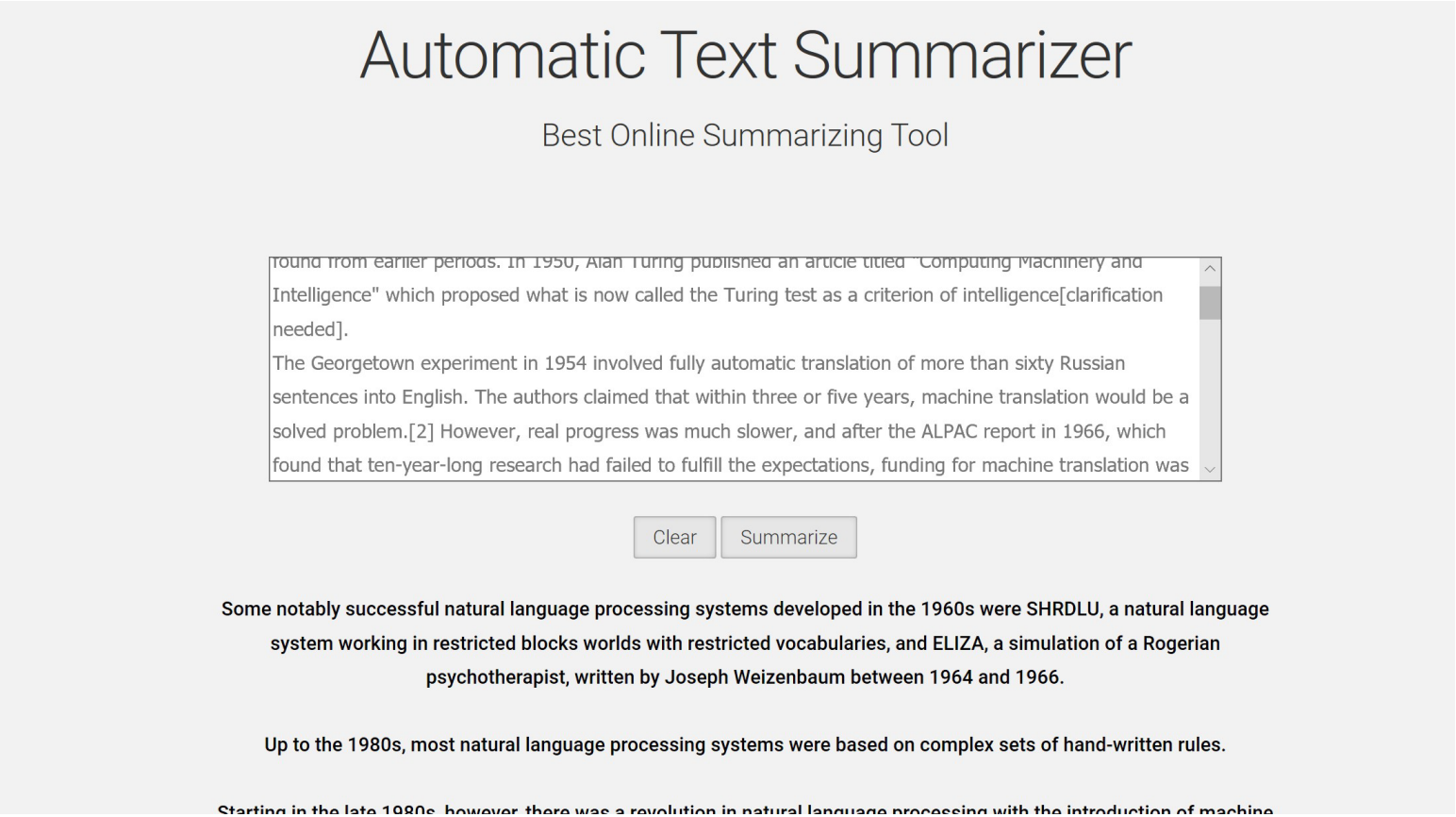
Spoken Dialog System

Siri.
Your wish is
its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

IBM Watson Developer Cloud

Summarization



Recap of Machine Learning

1. What are the similarities and differences between logistic regression and neural networks?
2. Given the function $f(x) = x^3$, what is the value of the slope at $x = 1$?
3. What is Stochastic Gradient Descent, and how is it functionally different to Batch Gradient Descent?
4. How does regularisation reduce overfitting? Explain this given the cost function with regularisation.

$$L_1 = (wx + b - y)^2 + \lambda|w|$$

$$L_2 = (wx + b - y)^2 + \lambda w^2$$

https://towardsdatascience.com/intuitions-on-l1-and-l2-regularisation-235f2db4c261

5. Share your impressions of using the tensorflow playground. What have you learned, what was interesting?

THE UNIVERSITY of ADELAIDE