

WHAT IS NLP?

Yogesh Kulkarni

July 25, 2022

Use case: Bank Call Center

Calling the Call Center

- ▶ Calling to an IVR (Integrated voice response)
- ▶ A prerecorded menu selection.
- ▶ “Please press 1 for Account Details, Please press 2 for . . . ”
- ▶ till it comes to your option.
- ▶ towards end, somewhere, given access to a person to talk to.

Boring? Annoying?



(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

Instead, how about typing/saying your query directly and getting the answer right away?

Solution

Chatbots

- ▶ Which problem of IVR it is solving?
- ▶ Advantages?
- ▶ Disadvantages?
- ▶ Gaining popularity ...
- ▶ Many platforms
- ▶ Companies in Pune?

The Giants are at it ...



(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

- ▶ Chatbots or QA systems, predominantly voice based,
- ▶ Underlying processing is primarily Natural Language Processing (NLP).
- ▶ You can have your own chatbot, specific to you!!
- ▶ NLP is the core skill needed.

Why so much popularity?

Chatbots are:

- ▶ Autonomous and Always Available
- ▶ Drive Conversation
- ▶ Able to handle millions of requests, scalable.

Its hard to master language, and thus NLP.

NLP is AI-complete

- ▶ “The most difficult problems in AI manifest themselves in human language phenomena.”
- ▶ Use of language is the touchstone of intelligent behavior.
- ▶ Test for Intelligence - Turing Test
- ▶ Alan Turing (1950) proposed a test of a machine's capability to perform human-like conversation.

Turing Test

A human judge engages in a natural language conversation with two other parties, one a human and the other a machine; if the judge cannot reliably tell which is which, then the machine is said to pass the test.



Early Conversational Programs

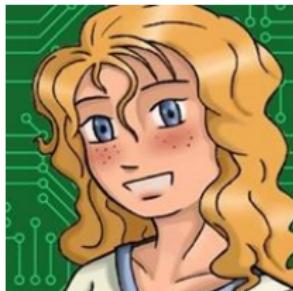
- ▶ ELIZA (by Joseph Weizenbaum), 1966.
- ▶ A psychotherapist, but NO real understanding;
- ▶ Simple pattern-matching to respond to user input to canned responses

```
(my ?x depresses me) (why does your ?x depress you)  
(life ?x) (why do you say it ?x)  
(I could ?x) (you could ?x)  
(because ?x) (that is a good reason)  
(?x) (tell me more)
```

```
>> Hello.  
How do you do. Please state your problem.  
>> Men are all alike.  
In what way ?  
>> They're always bugging us about something or other.  
I'm not sure I understand you fully.  
>> They're always bugging us about something or other.  
Can you think of a specific example ?  
>> Well, my boyfriend made me come here.  
Your boyfriend made you come here ?  
>> He says I'm depressed much of the time.  
I am sorry to hear that you are depressed.
```

Loebner Prize

- ▶ In 1990, Hugh Loebner started Turing Test competition
- ▶ \$100,000 will be awarded to the first bot that judges cannot distinguish from a real human in a Turing test that includes text, visual, and auditory input.
- ▶ Nobody has won the grand prize yet.
- ▶ 2016 (and 2013) year-wise top winner - Mitsuku.
<https://www.facebook.com/mitsukubot>



Why can't we win the Grand Prize? What are the challenges? Why Language is hard? What is Language?

What is Language?

Language Types

Natural Language

Winter und Sommer sind unterschiedlich, im Herbst
sind sie ähnlich und im Frühling sind sie wieder
ähnlich. Es gibt viele verschiedene Arten von Wetter.
Die Wettervorhersage ist eine Art von Wetterbericht,
der Ihnen sagt, was das Wetter morgen sein wird.
Sie können Wetterberichte auf dem Internet
oder in Zeitungen finden. Ein Wetterbericht kann
eine Temperaturangabe, eine Windrichtung und eine
Windstärke enthalten. Ein Wetterbericht kann auch
eine Aussicht auf Regen oder Schneefall haben.
Ein Wetterbericht kann auch eine Aussicht auf
Sonnenschein oder Regen haben.

(<http://expertenough.com/2392/german-language-hacks>)

日本語で

あゆせかいからくちいわおこなじき
冬は世界各地でさまざまなお祝いが行われる時期です。
ほんのいくつからいわきを挙げるだけでも、ハナカ、クリスマス、クワンザ、新年などさまざまなお祝いがあります。
かくさんかく文化によってその祝い方はさまざまですが、ほとんどのお祝いにはごちそうが欠かせません。

(http://www.transparent.com/learn-japanese/articles/dec_99.html)

Artificial Language

```
try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.get
            acWriter.println(cMessage);
            acWriter.flush();
        } catch (IOException e) {
            System.out.append(char c)
            queue.append(CharSequence s)
        }
    }
} catch (InterruptedException e) {
    System.out.printf(Locale l, S
}
```

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

```
def addS(x):
    return x*5

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

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

Differences?

Language, simplistically

- ▶ A vocabulary consists of a set of words
- ▶ A text is composed of a sequence of words from a vocabulary
- ▶ A language is constructed of a set of all possible texts



(<http://learnenglish.britishcouncil.org/en/vocabulary-games>)

THIS WEEK

Beyond the genome

Study of the expressive repertoire of a new healthy and disease human genome could provide crucial information that predict variation and disease.

The first complete healthy human genome has been published, providing a wealth of genetic information that could help predict individual variation and disease risk. The genome of a healthy person is a complex mixture of genes, regulatory elements and other DNA sequences that have been passed down through generations. By comparing this genome with others, researchers can identify specific genetic variations that may be associated with certain diseases or traits. This knowledge can help in the development of personalized medicine, where treatments are tailored to individual patients based on their genetic profile. The genome also provides insights into the evolution of humans and the relationship between different populations. The study of the genome can also help in understanding the biology of other organisms, such as plants and animals, by comparing their genomes with the human genome. Overall, the study of the genome is a powerful tool for advancing medical research and improving human health.

Editorials

THIS WEEK

Beyond the genome

Study of the expressive repertoire of a new healthy and disease human genome could provide crucial information that predict variation and disease.

The first complete healthy human genome has been published, providing a wealth of genetic information that could help predict individual variation and disease risk. The genome of a healthy person is a complex mixture of genes, regulatory elements and other DNA sequences that have been passed down through generations. By comparing this genome with others, researchers can identify specific genetic variations that may be associated with certain diseases or traits. This knowledge can help in the development of personalized medicine, where treatments are tailored to individual patients based on their genetic profile. The genome also provides insights into the evolution of humans and the relationship between different populations. The study of the genome can also help in understanding the biology of other organisms, such as plants and animals, by comparing their genomes with the human genome. Overall, the study of the genome is a powerful tool for advancing medical research and improving human health.



(<http://www.old-engl.sh/language.php>)

(http://www.nature.com/polopoly_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf)

NLP

- ▶ NLP is Natural Language Processing, ie processing Natural Langauge for some end-purpose in mind.
- ▶ Inspite of usage of Natural Language for thousands of years, why are we not able to process it well?

NLP Challenges

Paraphrasing

Paraphrasing: Different words/sentences express the same meaning

- ▶ Season of the year: Fall/Autumn
- ▶ Book delivery time
 - ▶ When will my book arrive?
 - ▶ When will I receive my book?

Ambiguity

Ambiguity: One word/sentence can have different meanings

- ▶ Fall
 - ▶ The third season of the year
 - ▶ Moving down towards the ground or towards a lower position
- ▶ The door is open
 - ▶ Expressing a fact
 - ▶ A request to close the door

Syntax and ambiguity

"I saw the man with a telescope."

- Who had the telescope?

Semantics

The astronomer loves the star.

- ▶ Star in the sky
- ▶ Celebrity



(<http://en.wikipedia.org/wiki/Star#/media/File:Starsinthesky.jpg>)



(<http://www.businessnewsdaily.com/2023-celebrity-hiring.html>)

NLP Applications

Grammar

Spell and Grammar Checking

- ▶ Checking spelling and grammar
- ▶ Suggesting alternatives for the errors



All Images News Videos Books More ▾ Search tools

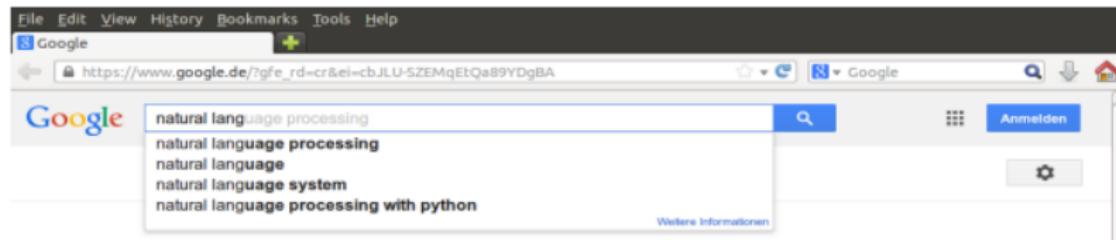
About 28.500.000 results (0,45 seconds)

Showing results for **natural** language processing
Search instead for **narural** language processing

Word Prediction

Word Prediction: Predicting the next word that is highly probable to be typed by the user

- ▶ Mobile typing
- ▶ Search Engines



Information Retrieval

Information Retrieval: Finding relevant information to the user's query

The screenshot shows a Google search results page for the query "panama papers". The search bar at the top contains the query. Below it, a navigation bar offers options: All (selected), Images, Shopping, News, Videos, More, and Search tools. A status message indicates "About 88.000.000 results (0,57 seconds)".

The first result is a news article from [sueddeutsche.de](http://www.sueddeutsche.de/panamapapers) titled "Datenleak Panama Papers - sueddeutsche.de". It includes a snippet of text: "Alle Details zu den Enthüllungen jetzt mit SZ Plus lesen Bleiben Sie informiert · Alle News zum Thema · Immer aktuell".

The second result is a link to "The Panama Papers · ICIJ" with the URL <https://panamapapers.icij.org/>. The snippet describes it as "Politicians, Criminals and the Rogue Industry That Hides Their Cash."

The third result is a link to "Panama Papers - Wikipedia, the free encyclopedia" with the URL https://en.wikipedia.org/wiki/Panama_Papers. The snippet states: "The Panama Papers are a leaked set of 11.5 million confidential documents that provide detailed information about more than 214,000 offshore companies ...".

A section titled "In the news" displays a thumbnail image of two men in suits and a news headline: "Panama Papers: Putin rejects corruption allegations - BBC News". The snippet below the headline reads: "President Putin has denied "any element of corruption" over the Panama Papers leaks, ...".

Below this, another news item is listed: "Panama Papers: David Cameron admits profiting from fund".

Text Categorization

Text Categorization: Assigning one (or more) pre-defined category to a text

The screenshot shows a PubMed search results page. At the top, there's a navigation bar with 'PubMed' and 'Advanced' search options. Below it, a search bar contains the term 'Abstract'. To the right of the search bar are 'Send to:' and 'Display Settings' buttons. The main content area displays an article abstract from 'Nature' (2014 Mar 20;507(7492):323-8). The abstract discusses the coupling of angiogenesis and osteogenesis in bone. It's authored by Kusumbe AA, Bamashmy S, and Adams RH. The abstract highlights that the mammalian skeletal system harbors a hierarchical system of mesenchymal stem cells, osteoprogenitors, and osteoblasts that sustain lifelong bone formation. Osteogenesis is indispensable for homeostatic renewal of bone as well as regenerative fracture healing, but these processes frequently decline in aging organisms, leading to loss of bone mass and increased fracture incidence. Evidence indicates that the growth of blood vessels in bone and osteogenesis are coupled, but relatively little is known about the underlying cellular and molecular mechanisms. Here we identify a new capillary subtype in the murine skeletal system with distinct morphological, molecular, and functional properties. These vessels are found in specific locations, mediate growth of the bone vasculature, generate distinct metabolic and molecular microenvironments, maintain perivascular osteoprogenitors and couple angiogenesis to osteogenesis. The abundance of these vessels and associated osteoprogenitors was strongly reduced in bone from aged animals, and pharmacological reversal of this decline allowed the restoration of bone mass.

MeSH Terms

Aging/metabolism
Aging/pathology
Animals
Blood Vessels/anatomy & histology
Blood Vessels/cytology
Blood Vessels/growth & development
Blood Vessels/physiology*
Bone and Bones/blood supply*
Bone and Bones/cytology
Endothelial Cells/metabolism
Hypoxia-Inducible Factor 1, alpha Subunit/metabolism
Male
Mice
Mice, Inbred C57BL
Neovascularization, Physiologic/physiology*
Osteoblasts/cytology
Osteoblasts/metabolism
Osteogenesis/physiology*
Oxygen/metabolism
Stem Cells/cytology
Stem Cells/metabolism

Text Categorization



Classify

Classify method: text url

Enter url to download and classify with:

uClassify!

Remove html

1. Sports (92.8 %)
2. Entertainment (4.8 %)
3. Men (0.7 %)

[Show all classifications >>](#)

Summarization

Summarization: Generating a short summary from one or more documents, sometimes based on a given query



This is a 7 sentence summary of <http://hpi.de/en/news/jahrgaenge/2015/des...>

Summary processing at low priority, upgrade to BOOST

Design Thinking Week: Students Improve the Daily Life Experience for People with Illiteracies

On the occasion of the World Literacy Day on September 8 more than 40 young innovators applied their Design Thinking skills in order to make life easier for these people.

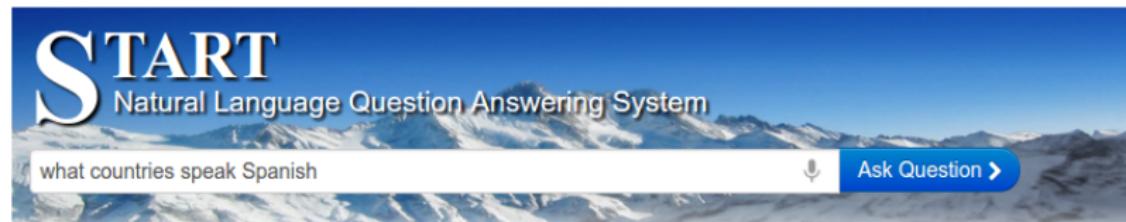
Here, the focus was especially on the possibilities of using digital technologies and computers to better the daily obstacles in life of the people concerned.

Under the guidance of the D-School's coaches the teams researched, developed and prototyped - and could present many versatile solutions in the end: e.g. one of the groups came up with an idea for a software program that lets internet browsers read texts, functions and links out loud so that people with reading problems can still use news sites or social networks like Facebook.

<http://smmry.com/>

Question answering

Question answering: Answering questions with a short answer



==> what countries speak Spanish

The language Spanish is spoken in Argentina, Aruba, Belize, Bolivia, Brazil, Canada, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Falkland Islands (Islas Malvinas), Gibraltar, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Martin, Sint Maarten, Spain, Switzerland, Trinidad and Tobago, United States, Uruguay, Venezuela, and Virgin Islands.

The language Castilian Spanish is spoken in Spain.

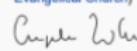
Question answering

Question answering: IBM Watson in Jeopardy



Information Extraction

Information Extraction: Extracting important concepts from texts and assigning them to slot in a certain template

 Merkel at the EPP Summit, March 2016	In office 17 November 1994 – 26 October 1998 Chancellor Helmut Kohl Preceded by Klaus Töpfer Succeeded by Jürgen Trittin Minister for Women and Youth In office 18 January 1991 – 17 November 1994 Chancellor Helmut Kohl Preceded by Ursula Lehr Succeeded by Claudia Nowotny Personal details Born Angela Dorothea Kasner 17 July 1954 (age 61) Hamburg, West Germany Political party Democratic Awakening (1989–1990) Christian Democratic Union (1990–present) Spouse(s) Ulrich Merkel (1977–1982) Joachim Sauer (1998–present) Alma mater Leipzig University Religion Lutheranism (within Evangelical Church) Signature 
---	---

Information Extraction

Information Extraction: Includes named-entity recognition

 lancet
a Medication Event Extraction System for Clinical Text

Project Home Downloads Wiki Issues Source
Summary People

Project Information

Started by 1 user
[Project feeds](#)

Code license
[GNU GPL v2](#)

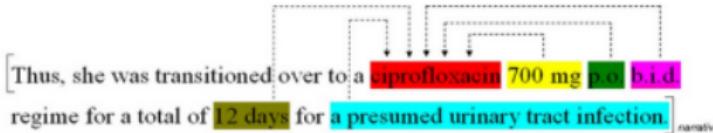
Labels
medication, extractor, lancet, discharge, summary, i2b2, NLP, challenge, 2009

 Members
lizuof...@gmail.com

Lancet is a supervised machine-learning system that automatically extracts medication events consisting of medication names and information pertaining to their prescribed use (dosage, mode, frequency, duration and reason) from lists or narrative text in medical discharge summaries.

Thus, she was transitioned over to a ciprofloxacin 700 mg p.o. b.i.d. regime for a total of 12 days for a presumed urinary tract infection

■ = medication ■ = dosage ■ = manner ■ = frequency ■ = duration ■ = reason



Machine Translation

Machine Translation: Translating a text from one language to another

The screenshot shows the Google Translate interface. At the top, there's a search bar with the word "Translate". Below it, a toolbar has buttons for "German", "Portuguese", "Spanish", "Detect language", and a dropdown menu. To the right of the toolbar are buttons for "English", "Portuguese", "German", and a "Translate" button. The main area contains two text boxes. The left text box contains the German sentence: "Die Lehre am Hasso-Plattner-Institut richtet sich an begabte junge Leute, die praxisnah zu IT-Ingenieuren ausgebildet werden wollen." The right text box contains the Portuguese translation: "Ensinar no Instituto Hasso Plattner é destinado a jovens talentosos que querem ser treinados para a prática de engenheiros de TI." Both text boxes have edit icons (trash, copy/paste, etc.) at the bottom.

Sentiment Analysis

Sentiment Analysis: Identifying sentiments and opinions stated in a text

Customer Reviews

Speech and Language Processing, 2nd Edition



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



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

Sentiment Analysis

Restaurant/hotel recommendation, Product reviews

Bodo's Bagels

Find Local, cheap dinner, Mac's Near Charlottesville, VA

Yelp logo

5 - Bagels, Breakfast & Brunch, Sandwiches

4.5 stars, 188 reviews [See Details](#)

Write a Review Add Photo Share Bookmarks

Address: 1418 Emmet St, Charlottesville, VA 22903 Get Directions Call: 434-295-5888 Manage the business [View Website](#)

Map showing location on Emmet Street, Charlottesville, VA.

Three images of bagels: one plain, one with cream cheese, and one with meat.

Turkey with lettuce and pickles and onions - by Zach R.

See all 30 reviews

"Almost any combination of bagel, cream cheese or spread or sandwich you could dream of you can find at Bodo's." in 30 reviews.

\$8.40 Cream Cheese

"A few favorite items would include the Everything bagel with the Deli Egg which has a tasty meaty center encased in steaming hot eggs." in 4 reviews

There's a reason why Bodo's has been in business since well before I was born.

Recommended Reviews

Sort by Highest Rated [Search reviews](#) English (186)

New York City > Hotels > Flights > Vacation Rentals > Restaurants > Things to Do > Best of 2013 Your Friends More Write a Review

New York City, New York, United States

What are you looking for? Search

Hilton Times Square

4.5 stars, 4,313 reviews #78 of 457 HOTELS in New York City Certificate of Excellence

+1 866-213-3621 Hotel deals Hotel website 8234 West 42nd Street, New York City, NY 10036

Special Offer [Request Special Offer](#)

PriceFinder Enter dates for best prices Check In Check Out Check Availability

Book on [tripadvisor](#) or compare prices there up to 200 sites including:

Booking.com Expedia Expedia

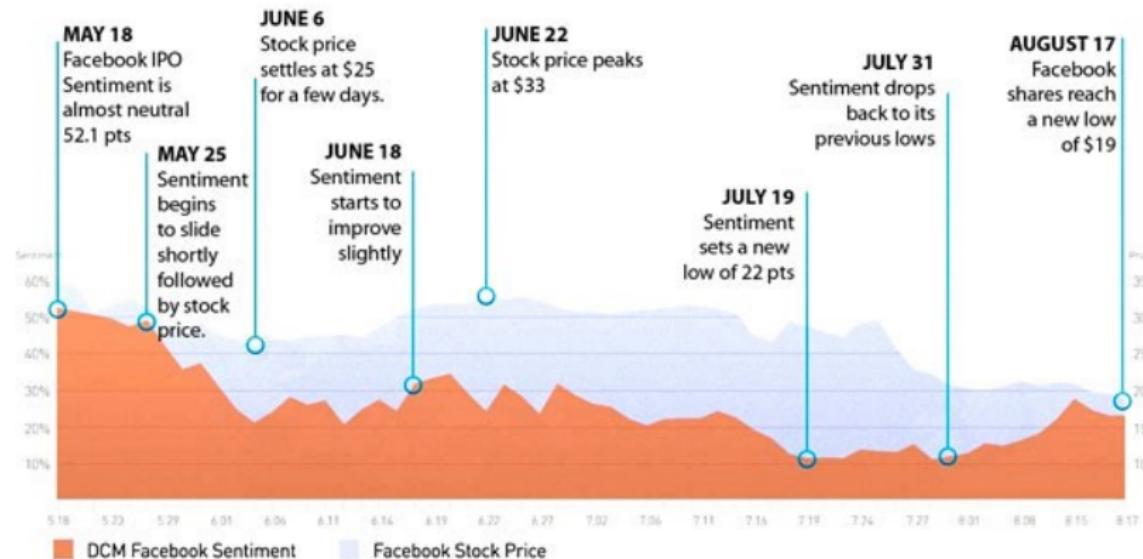
Applebee's Hilton Money Exchange

Reviews (4,919) Photos (1,654) Location Amenities GAA (129) Room Tips (1,085) Write a Review Add Photo

4,919 Reviews from our TripAdvisor Community

Sentiment Analysis

Text analytics in financial services



NLP in today's time

Trends:

- ▶ An enormous amount of information is now available in machine readable form as natural language text (newspapers, web pages, medical records, financial filings, product reviews, discussion forums, etc.)
- ▶ Conversational agents are becoming an important form of human-computer communication
- ▶ Much of human-human interaction is now mediated by computers via social media

Collectively, this means that copious data is available to be used in the development of NLP systems.

Level of difficulties

- ▶ Easy (mostly solved)
 - ▶ Spell and grammar checking
 - ▶ Some text categorization tasks
 - ▶ Some named-entity recognition tasks
- ▶ Intermediate (good progress)
 - ▶ Information retrieval
 - ▶ Sentiment analysis
 - ▶ Machine translation
 - ▶ Information extraction
- ▶ Difficult (still hard)
 - ▶ Question answering
 - ▶ Summarization
 - ▶ Dialog systems

NLP Activities: How to process text?

Sentence splitting

Sentence splitting: Splitting a text into sentences

11 Sentences (= "T-" or "Terminable" units only if independent clauses are punctuated as separate sentences, e.g. "I came and he went"-->"I came. And he went.")
Average 23.55 words (SD=12.10)

OBJECTIVES: To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinoma.

METHODS: Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Morphology features and vascularization perfusion on 3D ultrasound were evaluated.

Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c erbB-2 and p53 expression, and microvessel density (MVD) were determined.

Correlations of 3D ultrasound features and prognostic factors were analysed.

RESULTS: The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size ($P = 0.014$), a lower histological grade ($P = 0.009$) and positive ER or PR expression status ($P = 0.001, 0.044$).

The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours.

The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer.

The increased intra-tumour vascularization index (VI, the mean tumour vascularity) reflected a higher histological grade ($P = 0.025$) and had a positive correlation with MVD ($r = 0.530$, $P = 0.001$).

CONCLUSIONS: The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

KEY POINTS: • Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer. • The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis. • The increased intra-tumour vascularization index reflects a higher histological grade. • The intra-tumour vascularization index is positively correlated with microvessel density.

Tokenization

Tokenization

- ▶ Process of breaking a stream of text up into tokens (= words, phrases, symbols, or other meaningful elements)
- ▶ Typically performed at the “word” level
- ▶ Not easy: Hewlett-Packard, U.S.A., in some languages there is no “space” between words!

Stemming

Stemming

- ▶ Reduces similar words to a given “stem”
- ▶ E.g. detects, detected, detecting, detect : detect (stem).
- ▶ Usually set of rules for suffix stripping
- ▶ Most popular for English: Porter's Algorithm
- ▶ 36% reduction in indexing vocabulary (English)
- ▶ Linguistic correctness of resulting stems not necessary (sensitivities : sensit)

Lemmatization

Lemmatization

- ▶ Uses a vocabulary and full morphological analysis of words
- ▶ Aims to remove inflectional endings only
- ▶ Return the base or dictionary form of a word, which is known as the lemma.
- ▶ E.g. saw : see,
been, was : be

Part-of-speech tagging

Part-of-speech tagging: Assigning a syntactic tag to each word in a sentence

Stanford Parser

Please enter a sentence to be parsed:

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Language: English ▾

Sample Sentence

Parse

Your query

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Tagging

Surgical/NNP resection/NN specimens/NNS of/IN 85/CD invasive/JJ
ductal/JJ carcinomas/NNS of/IN 85/CD women/NNS who/WP had/VBD
undergone/VBN 3D/CD ultrasound/NN were/VBD included/VBN ./.

<http://nlp.stanford.edu:8080/corenlp/>

Parsing

Parsing: Building the syntactic tree of a sentence

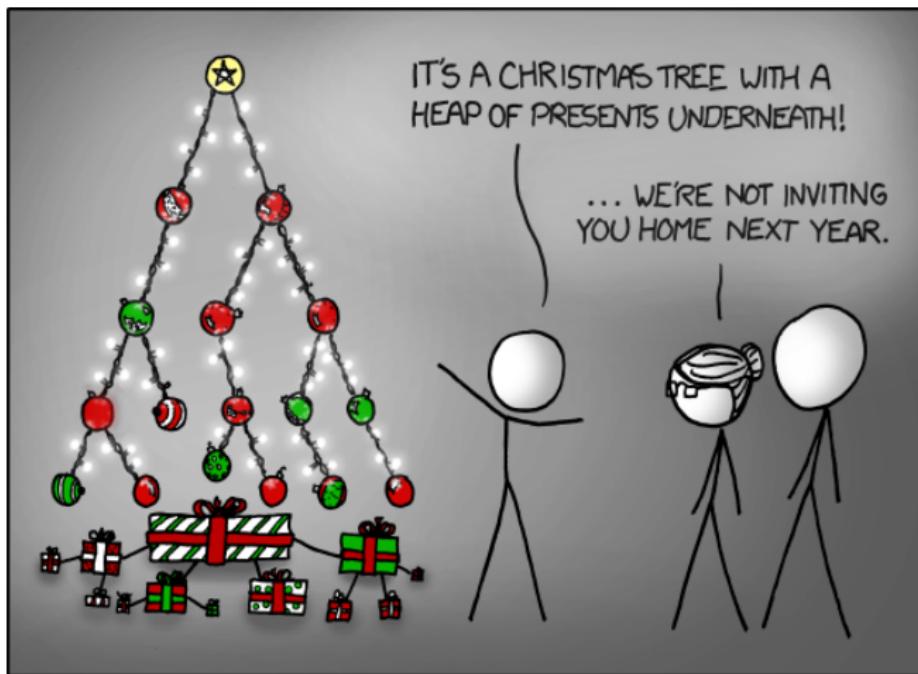
Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound)))))))))))
        (VP (VBD were)
          (VP (VBN included))))
      (. .)))
```

<http://nlp.stanford.edu:8080/corenlp/>

Parsing

$((DaimlerChryslershares)_{NP}(rose(threeeights)_{NUMP}(to22)_{PP-NUM})_{VP})_S$



Syntax Tree

Syntax: Sample English grammar

$S \rightarrow NP VP$

$S \rightarrow Aux NP VP$

$S \rightarrow VP$

$NP \rightarrow Pronoun$

$NP \rightarrow Proper-Noun$

$NP \rightarrow Det Nominal$

$Nominal \rightarrow Noun$

$Nominal \rightarrow Nominal Noun$

$Nominal \rightarrow Nominal PP$

$VP \rightarrow Verb$

$VP \rightarrow Verb NP$

$VP \rightarrow Verb NP PP$

$VP \rightarrow Verb PP$

$VP \rightarrow VP PP$

$PP \rightarrow Preposition NP$

$Det \rightarrow that | this | a$

$Noun \rightarrow book | flight | meal | money$

$Verb \rightarrow book | include | prefer$

$Pronoun \rightarrow I | she | me$

$Proper-Noun \rightarrow Houston | TWA$

$Aux \rightarrow does$

$Preposition \rightarrow from | to | on | near | through$

Named-entity recognition

Named-entity recognition: Identifying pre-defined entity types in a sentence

bio2rdf Annotate

HIGHLIGHT

- All None
- Anatomy
- Chemical
- Genes and Proteins
- Cellular Components
- Molecular Functions
- Biological Processes
- Ambiguous

In **Duchenne muscular dystrophy** (DMD), the **infiltration** of **skeletal muscle** by immune **cells** aggravates disease, yet the precise mechanisms behind these inflammatory responses remain poorly understood. Chemokines, or chemoattractants, are considered essential regulators of **inflammatory cells** to the **tissue**. We assayed chemokine and chemoattractant receptor expression in **CMD** muscle biopsies ($n = 9$, average age 7 years) using immunohistochemistry, immunofluorescence, and *in situ* hybridization. CXCL1, CXCL2, CXCL3, CXCL8, and CXCL11, absent from normal **muscle** fibers, were induced in **DMD** myofibers. CXCL1, CXCL2, and the ligand-receptor couple CXCL2-CCR2 were upregulated on the **blood vessel** endothelium of **DMD** patients. CCR1 (+) **macrophages** expressed high levels of CXCL8, CXCL2, and COL5. Our data suggest a possible beneficial role for CXCL1/2 ligands in managing **muscle fiber** damage control and **tissue** regeneration. Upregulation of **endothelial** chemoattractant receptors and CXCL8, CXCL2, and COL5 expression by cytotoxic **macrophages** may regulate myofiber **regrowth**.

Lead text Export ▾

Annotated 46 concept occurrences in 0.173s.

Now to focus? Take the tour ▾

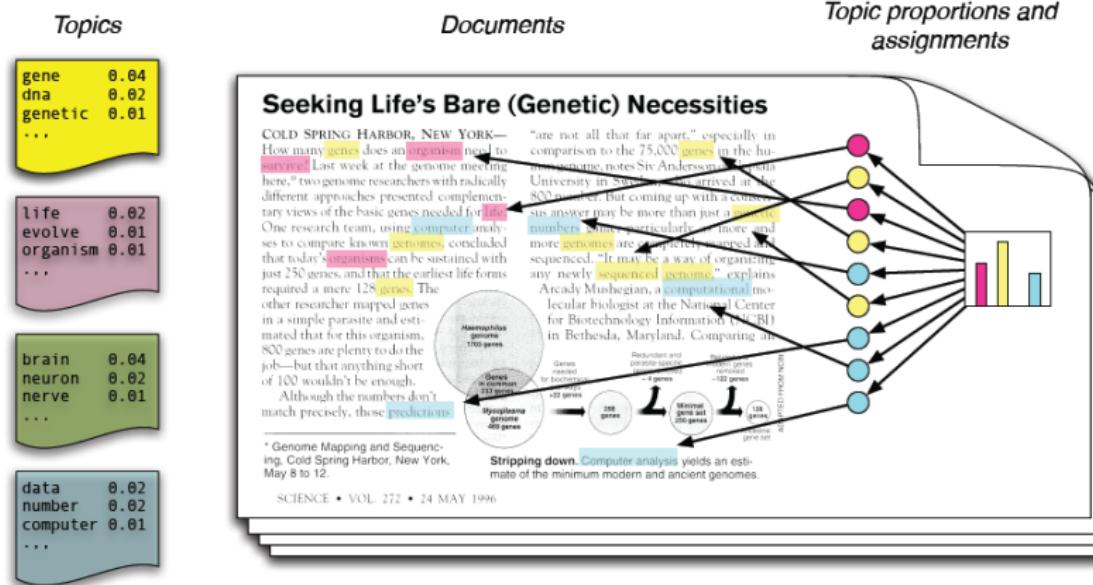
Expand All ▾ Collapse All ▾ Toggle All ▾

Concept Tree

- Anatomy (12)
 - Disorders (4)
 - DMD (1)
 - Duchenne muscular dystrophy (1)
 - Infiltration (1)
 - Inflammatory responses (1)
- Chemicals (2)
- Genes and Proteins (11)
- Cellular Components (3)
- Molecular Functions (1)
- Biological Processes (9)

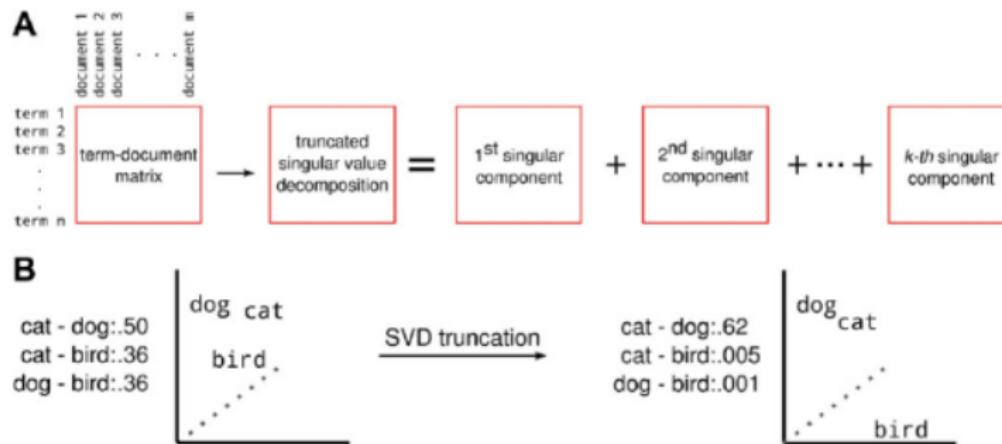
Topic modelings

Topic modeling: Identifying structures in the text corpus



Word embeddings

Word embeddings: Compute a vector representing the distributed representation for every word



What next?

- ▶ Coursera : Dr Radev's NLP course
(<https://www.coursera.org/learn/natural-language-processing>)
 - ▶ Course: Deep NLP By Richard Socher (Stanford)
 - ▶ Book: Natural Language Processing with Python



NLP Opportunities



Speech
Transcription



Neural Machine
Translation (NMT)



Chatbots



Q&A



Text
Summarization



Image
Captioning



Video
Captioning

(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

References

Many publicly available resources have been refereed for making this presentation. Some of the notable ones are:

- ▶ Introduction to Natural Language Processing - Dr. Mariana Neves, SoSe 2016
- ▶ Machine Learning for Natural Language Processing - Traian Rebedea, Stefan Ruseti - LeMAS 2016 - Summer School
- ▶ CSC 594 Topics in AI - Natural Language Processing - De Paul
- ▶ Deep Learning for Natural Language Processing - Sihem Romdhani
- ▶ Notebooks and Material @
https://github.com/rouseguy/DeepLearningNLP_Py

Thanks ... yogeshkulkarni@yahoo.com