Lets Learn NLP

Yogesh Haribhau Kulkarni

YHK

Outline

About Me

Yogesh Haribhau Kulkarni

Bio:

- 20+ years in CAD/Engineering software development
- Got Bachelors, Masters and Doctoral degrees in Mechanical Engineering (specialization: Geometric Modeling Algorithms).
- Currently doing Coaching in fields such as Data Science, Artificial Intelligence Machine-Deep Learning (ML/DL) and Natural Language Processing (NLP).
- Feel free to follow me at:
  - Github (github.com/yogeshhk)
  - LinkedIn (www.linkedin.com/in/yogeshkulkarni/)
  - Medium (yogeshharibhaukulkarni.medium.com)
  - Send email to yogeshkulkarni at yahoo dot com

Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.

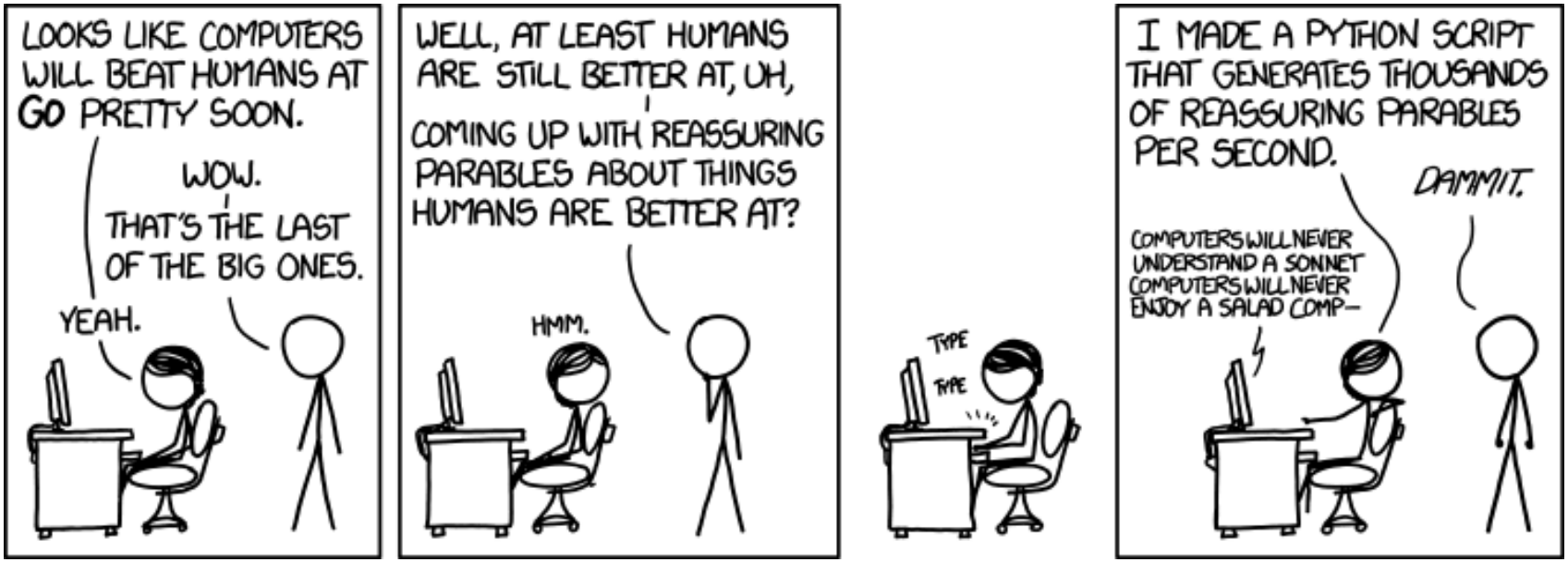
YHK

Natural Language Processing: An Introduction

YHK

Background: What is Language?

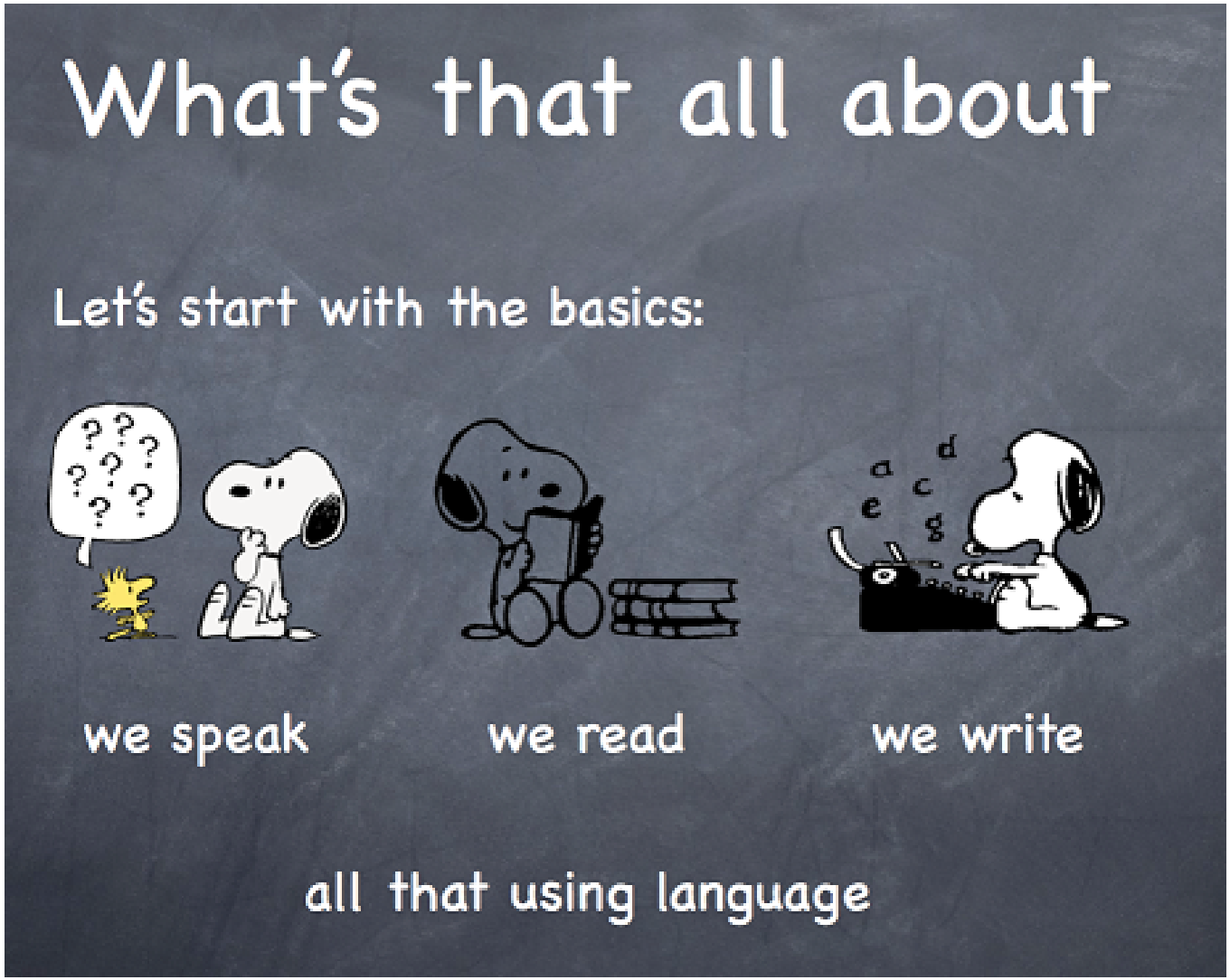
xkcd



(Parables: verse/short stories with morals)

YHK

What's a Language?



What's a Language?

But that's not all...

- we think of the world around us
- we dream
- we make decisions and plans
- all in natural language, i.e. in words

YHK

Words == Meaning?

When you read the word, say, “Crab”, what does this mean to you?
Bunch of letters? or something else?

YHK

Words == Meaning?

/krœb/ **Crab**

Is the symbol representative of its meaning?

Words == Meaning?

Now, which of these you can understand?

sound	symbol	sight
/krœb/	**Crab**	

Or is it just a mapping in a mental lexicon/vocabulary/dictionary (word, picture, understanding)?

YHK

Arbitrariness of Language Symbols

sound	symbol	sight
/kávouras/	**κάβουρας**	

These symbols are arbitrary. Words don't embed knowledge by themselves.
But words are the only thing we have for processing!
Verbal and written communication is primarily through words.

YHK

Why Analyze Language?

Words/Language is:

- Main channel of communication
- Vehicle for knowledge acquisition
- Repository of human knowledge

Language technologies process human language automatically:

- Hand-held devices: predictive text, handwriting recognition
- Web search engines: access information in text
- Virtual assistants: natural human-machine interfaces
- Knowledge extraction: unlock information from documents

YHK

Discussion Questions

- Which language systems have you recently used?
- What is impressive about these systems?
  - Imagine you stepped out of a time machine from the 1970s or 1970s
- How could these systems be improved?
- What limitations do current systems have?

YHK

Introduction: What is NLP?

Motivation: Bank Call Center Use Case

Traditional IVR (Interactive Voice Response):

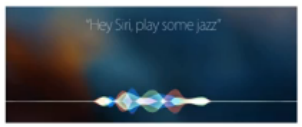
- Navigate through endless menus
- "Press 1 for Account Details, Press 2 for..."
- Frustrating user experience

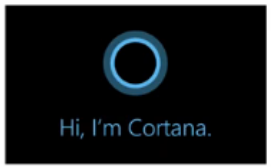
Modern Solution: Conversational AI powered by NLP

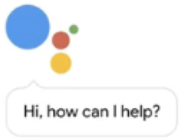
- Natural dialogue with chatbots
- Direct query answering
- 24/7 availability, scalable to millions

YHK

The AI Revolution in Language







Major Players in Conversational AI:

- Voice Assistants: Siri, Alexa, Google Assistant
- Chatbots: ChatGPT, Claude, Gemini
- Enterprise: Customer service, virtual agents
- All powered by Natural Language Processing

YHK

What is Natural Language Processing?

Definition and Scope

Natural Language Processing (NLP):

- Natural Language: Languages spoken by humans (English, French, German, etc.)
- As opposed to Formal Languages (Python, Java, C++, etc.)
- NLP: Computer applications that automatically analyze, understand, and generate natural language
- Interdisciplinary field: Computer Science + Linguistics + AI + Statistics

YHK

Formal vs. Natural Languages

Formal Languages

- Strict, unchanging rules
- Application specific
- Literal meaning
- No ambiguity
- Parsable by regular expressions
- Inflexible vocabulary

Natural Languages

- Flexible, evolving
- Domain-independent
- Redundant and verbose
- Highly ambiguous
- Difficult to parse
- Very flexible
- Context-dependent

YHK

Language Components

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

日本語で
冬は世界各地でさまざまなお祝いが行われる時期です。ほんのいくつか例を挙げるだけでも、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。各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)

Language = Words + Rules + Exceptions + Context

- Dictionary (Vocabulary): Set of words in the language
- Grammar: Rules describing what is allowable
- Semantics: Meaning of words and sentences
- Pragmatics: Context and social use

YHK

What is NLP?

Natual Language Processing

- Natural Language: languages spoken by people (English, French, German, etc.) as opposed to artificial languages, also called as Formal, (C++, Java, Python, etc.) built for computer manipulation
- Natural Language Processing: computer applications that automatically analyze natural language

YHK

What is NLP?

- Language = Words + Rules + Exceptions + ...
- Formally: dictionary (vocabulary) + grammar + ...
- Dictionary : set of words defined in the language (static or dynamic)
- Grammar: set of rules which describe what is allowable in a language

YHK

NLP Challenges: Paraphrasing

Paraphrasing: Different words/sentences express the same meaning

- Synonymy: Fall/Autumn
- Book delivery questions:
  - When will my book arrive?
  - When will I receive my book?
  - What's the delivery time for my book?
  - How long until my book gets here?

All mean the same thing but use different words and structures!

YHK

NLP Challenges: Ambiguity

Ambiguity: One word/sentence can have different meanings

- Lexical Ambiguity - "Fall":
  - The third season of the year
  - Moving down towards the ground
- Pragmatic Ambiguity - "The door is open":
  - Expressing a fact
  - A request to close the door
- Syntactic Ambiguity: "I saw the man with a telescope"
  - Who had the telescope?

YHK

NLP Challenges: Semantic Complexity

"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)

Context is crucial for disambiguation!

YHK

Why is NLP Hard?

Core Challenges:

- Ambiguity: Many ways to represent the same thing
- High dimensionality and sparsity: Many rare words
- Out-of-sample generalization: New words and sentences constantly
- Order and context matter: "Dog bites man" vs "Man bites dog"
- Implicit knowledge: Cultural context, world knowledge
- Non-compositionality: "kick the bucket" $\neq$ kick + bucket

YHK

The Richness of Natural Language

Language serves:

- Basic needs and lofty aspirations
- Technical know-how and flights of fantasy
- Ideas shared across distance and time

Examples (think of processing these!):

- "Overhead the day drives level and grey, hiding the sun by a flight of grey spears." (William Faulkner)
- "When using the toaster please ensure that the exhaust fan is turned on." (sign)
- "Iraqi Head Seeks Arms" (spoof headline)
- "Twas brillig, and the slithy toves did gyre and gimble in the wabe" (Lewis Carroll, Jabberwocky)
- "There are two ways to do this, AFAIK :smile:" (internet discussion)

YHK

Understanding Questions: Beyond Keywords

- What is the question really asking? (different from keyword search)
- Beyond finding candidate passages; choose the right one
- Example: "What is the fastest automobile in the world?"
  - A1: "...will stretch Volkswagen's lead in the world's fastest growing vehicle market..."
  - A2: "...the Jaguar XJ220 is the dearest (£415,000), fastest (217mph) and most sought after car in the world."
- Requires semantic understanding and reasoning
- What if answers require aggregation across sources?

YHK

The Challenge: Human Language is Messy

If humans may not follow other humans, what hope for machines?
But we can attempt! Need to study language systematically.

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth, UIUC)

YHK

NLP is AI

Can Machine understand the way humans do?

YHK

Can Machine Answer?

(ENGLAND, June, 1989) - Christopher Robin is alive and well. He lives in England. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book. He made up a fairy tale land where Chris lived. His friends were animals. There was a bear called Winnie the Pooh. There was also an owl and a young pig, called a piglet. All the animals were stuffed toys that Chris owned. Mr. Robin made them come to life with his words. The places in the story were all near Cotchfield Farm. Winnie the Pooh was written in 1925. Children still love to read about Christopher Robin and his animal friends. Most people don't know he is a real person who is grown now. He has written two books of his own. They tell what it is like to be famous.

1. Who is Christopher Robin?
2. When was Winnie the Pooh written?
3. What did Mr. Robin do when Chris was three years old?
4. Where did young Chris live?
5. Why did Chris write two books of his own?

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth University of Illinois, Urbana-Champaign)

YHK

Understanding Questions

- What is the question asking? (different from Googling)
- Beyond finding candidate passages; choose the right one.
- Say, Q: What is the fastest automobile in the world?
- A1: . . . will stretch Volkswagen's lead in the world's fastest growing vehicle market. Demand for cars is expected to soar . . .
- A2: . . . the Jaguar XJ220 is the dearest (415,000 pounds), fastest (217mph) and most sought after car in the world.
- And, what if the answers require aggregation

YHK

Not So Easy

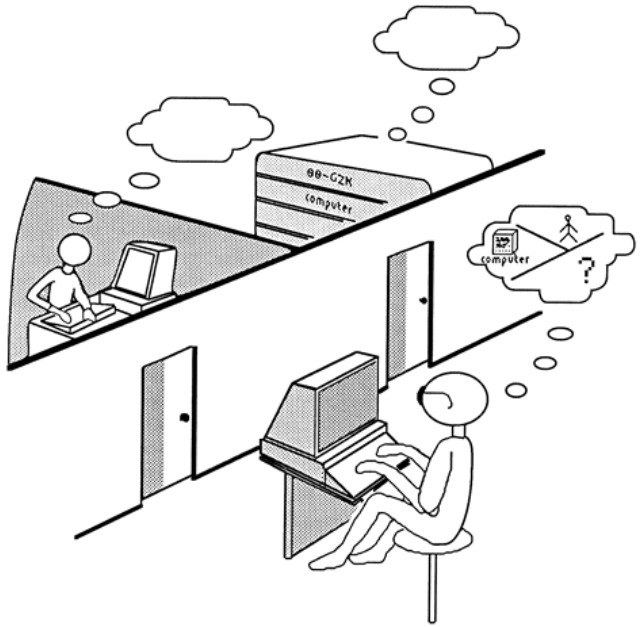
Humans may not follow other humans, then what for the machines. But still we can attempt.
Need to study the language well!!!

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth University of Illinois, Urbana-Champaign)

YHK

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
- The Turing Test (Alan Turing, 1950)



A human judge engages in natural language conversation with two parties (one human, one machine). If the judge cannot reliably distinguish them, the machine passes the test.

YHK

History of NLP

History of NLP: Early Foundations (1950s-1960s)

Antiquity

What is language?

- Plato (~400 BCE)
  - *Cratylus* – dialogue about language
  - Word mappings are innate, not learned
  - Words are intrinsically related to their meanings
- Aristotle (~350 BCE)
  - Language is deductive, like physics
  - Abstracted impressions

Plato[1]

Aristotle[2]

12

1: https://www.britannica.com/biography/Plato 2: https://www.britannica.com/biography/Aristotle 3: (Tomasello, 2010) 4: (Miner, 2012)

- 1950: Turing Test proposed
- 1954: Georgetown-IBM experiment (machine translation)
- 1960s: ELIZA, symbolic AI approaches
- Rule-based and deductive language systems

(Ref: Text Mining - Jeff Shaul)

YHK

Early Conversational Programs

ELIZA (Joseph Weizenbaum, 1966)

- Simulated a psychotherapist
- Simple pattern-matching with canned responses
- No real understanding

(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)

YHK

History of NLP: Linguistic Era (1970s-1980s)

Pre-1950s

Cataloguing

- Thomas Hyde, 1674[3]
  - Cataloguing by classification
  - Before Hyde, catalogued by author or title
- Melvil Dewey, 1876[4]
  - Index card catalog
  - Dewey decimal classification
- Claude Shannon, 1948[6]
  - *A Mathematical Theory of Communication*
  - "father of information theory"
  - Source > transmitter > channel > receiver > destination

Thomas Hyde[1]

Melvil Dewey[2]

Claude Shannon[5]

1: http://www.mindsacollection.org/lockes-library/ 2: http://alchetron.com/Melvil-Dewey-1187480-W 3: (Marshall, 2012) 4: (Satija, 2013)
5: http://www.newyorker.com/tech/elements/claude-shannon-the-father-of-the-information-age-turns-1100100 6: (Shannon, 1948)

13

- Chomskyan linguistics influence
- Context-free grammars
- Knowledge-based systems
- First AI winter (mid-1970s)

(Ref: Text Mining - Jeff Shaul)

YHK

History of NLP: Statistical Revolution (1990s)

1950s

Automatic abstracting

- H.P. Luhn, 1958[1,2]
  - *The Automatic Creation of Literature Abstracts*
  - Pioneer of auto-abstracting for business intelligence
  - Used a room-sized computer: IBM 704
  - Statistical method for automatic abstract generation of technical articles
- Noam Chomsky, 1959[4]
  - *Syntactic Structures*
  - "generative grammar"
  - Rule-based descriptions of syntactic structures
  - Extremely influential in computational linguistics

H.P. Luhn[3]

Noam Chomsky[5]

14

1: (Luhn, 1958a) 2: (Luhn, 1958b) 3: https://www.asist.org/wp-content/uploads/2015/01/Hans_Luhn.png 4: (Chomsky, 1959) 5: http://radioopensource.org/vietnam-worst-war/

- Shift from rule-based to statistical methods
- Hidden Markov Models (HMMs)
- N-gram language models
- Machine learning emerges

(Ref: Text Mining - Jeff Shaul)

YHK

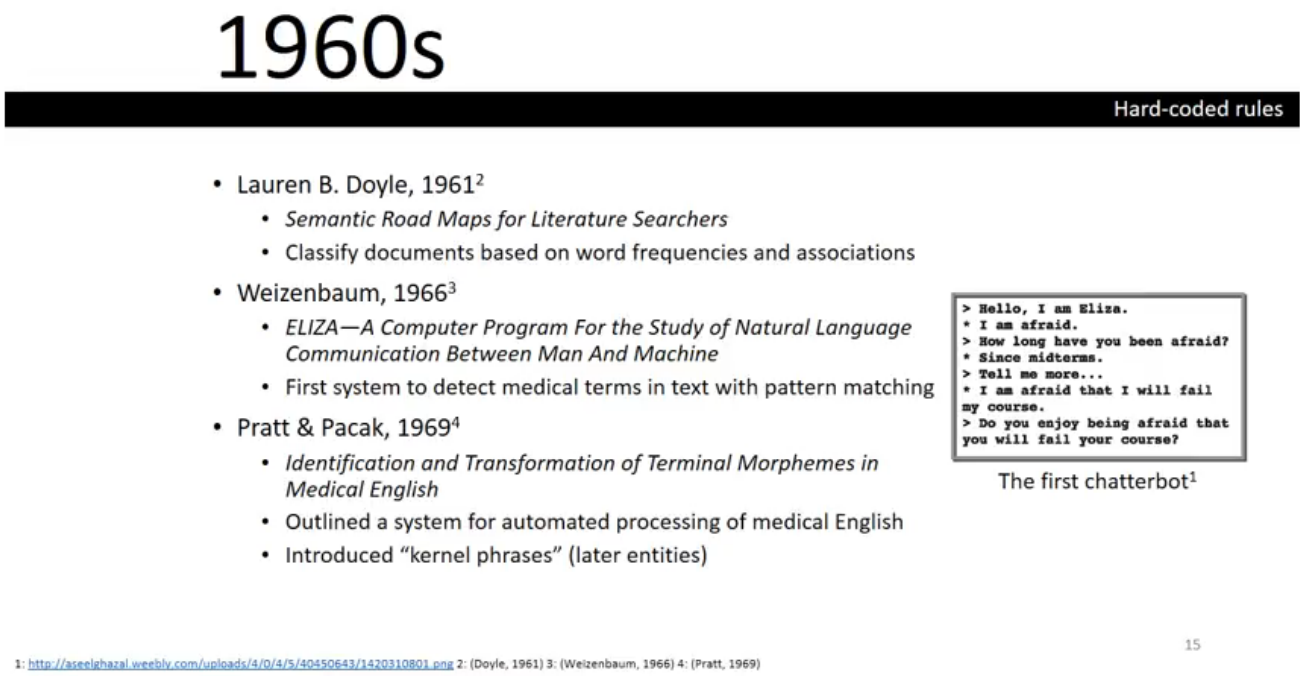
The Loebner Prize

- Started in 1990 by Hugh Loebner
- Annual Turing Test competition
- $100,000 for first bot indistinguishable from human
- Nobody has won the grand prize yet
- Recent winners: Mitsuku (2013, 2016-2019)

Why can't we win? Understanding language challenges is key.

YHK

History of NLP: Machine Learning Era (2000s)



- Supervised learning dominance
- Support Vector Machines (SVMs)
- Conditional Random Fields (CRFs)
- Large annotated corpora

(Ref: Text Mining - Jeff Shaul)

YHK

History of NLP: Deep Learning Revolution (2010s)

Early 2000s

Tools, tools, tools

- Friedman *et al*, 2001[3]
  - *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.*
  - First widespread/successful tool for extracting structured information from genomic literature
- Collins & Duffy, 2002[4]
  - *New ranking algorithms for parsing and tagging*
  - Improved use of decision trees in text mining
- Hotho *et al*, 2003[5]
  - *Wordnet improves Text Document Clustering*
  - Formal use of domain ontologies/lexicons
- Müller *et al*, 2004[6]
  - *Textpresso: An ontology-based information retrieval and extraction system for biological literature*
  - Robust literature-based discovery system

Hans-Michael Müller[1]

Andreas Hotho[2]

1: http://www.jbc.org/content/290/45/27015/suppl/DCAuthor_profile_HM 2: http://www.dmir.uni-wuerzburg.de/staff/hotho/
3: (Friedman, 2001) 4: (Collins, 2002) 5: (Hotho, 2003) 6: (Müller, 2004)

21

- Word embeddings: Word2Vec (2013), GloVe (2014)
- Recurrent Neural Networks (RNNs, LSTMs)
- Attention mechanisms
- Sequence-to-sequence models

(Ref: Text Mining - Jeff Shaul)

YHK

History of NLP: Transformer Era (2017-Present)

2010s

Supervision is for kids

- Collobert *et al*, 2011[3]
  - *Natural Language Processing (Almost) from Scratch*
  - Unified and unsupervised neural network framework for all tasks of NLP
- Das *et al*, 2011[4]
  - *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*
  - Widely applicable to languages without established ontologies/lexicons
- Zhang *et al*, 2013[5]
  - *Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts*
  - Stepwise approach to NER without any training data or heuristics

Ronan Collobert[1]

Dipanjan Das[2]

1: http://ronan.collobert.com/ 2: http://research.google.com/pubs/DipanjanDas.html
3: (Collobert, 2011) 4: (Das, 2011) 5: (Zhang, 2013) 6: (Zhang, 2016)

23

- 2017: "Attention is All You Need" - Transformers
- 2018: BERT, GPT-1
- 2019-2020: GPT-2, GPT-3, T5
- 2022-2024: ChatGPT, GPT-4, Claude, Gemini, LLaMA
- 2025: Multimodal models, reasoning capabilities

(Ref: Text Mining - Jeff Shaul)

YHK

Why NLP Now? The Perfect Storm

- **Data**: Massive digitized human knowledge
  - Web text, social media, books, articles
  - Twitter firehose of current events
- **Compute**: Fast GPUs and TPUs
  - Cloud computing enables scaling
  - Parallel processing capabilities
- **Algorithms**: Breakthrough architectures
  - Transformers and attention mechanisms
  - Self-supervised learning
  - Transfer learning and fine-tuning

YHK

Leveraging Big Data in NLP

- Examples are easier to create than rules
- Rules and logic miss frequency and language dynamics
- More data enables better machine learning
- Relevance is in the long tail
- Knowledge engineering doesn't scale
- Modern methodologies are data-driven and stochastic

Question: What do computers prefer - numbers or words?

YHK

NLP Approaches: Past & Present

Main Approaches in NLP

1. **Rule-based methods**
   - Regular expressions, context-free grammars
   - Hand-crafted linguistic rules
   - Example: Pattern matching for entity extraction
2. **Machine learning**
   - Feature engineering + classical ML
   - Linear classifiers, CRFs, SVMs
   - Example: Named Entity Recognition with CRF
3. **Deep Learning**
   - End-to-end neural networks
   - RNNs, LSTMs, Transformers
   - Example: BERT for text classification
4. **Large Language Models (2020s)**
   - Foundation models with billions of parameters
   - Few-shot and zero-shot learning
   - Example: GPT-4, Claude for question answering

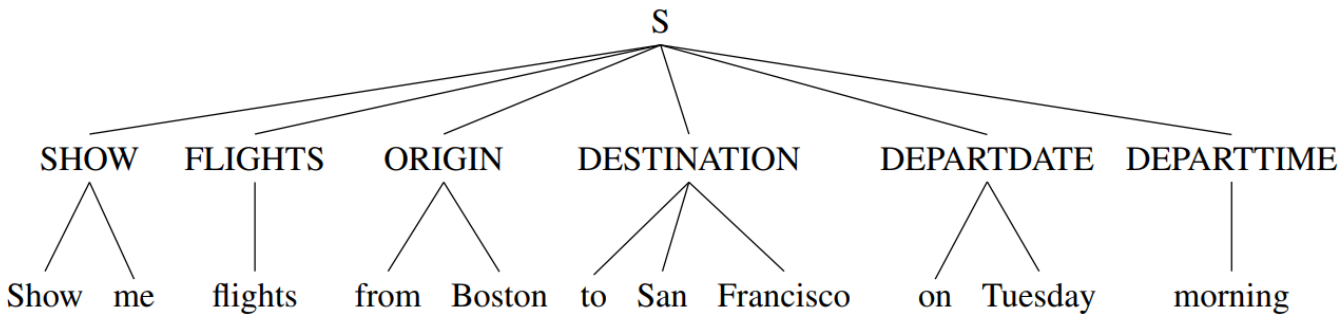
YHK

Example: Rule-based Approach

Task: Find slots (City, date, etc) from sentence
Input: "Show me flights from Boston to San Francisco on Tuesday"
Context-free grammar rules:

- SHOW: show me — i want — can i see
- FLIGHTS: (a) flight — flights
- ORIGIN: from CITY
- DESTINATION: to CITY
- CITY: Boston — San Francisco — Denver — Washington
- DAY: Monday — Tuesday — Wednesday — ...

Parse sentence and apply annotations:



(Ref: https://web.stanford.edu/ jurafsky/slp3/)

Example: Machine Learning Approach

CRF (Conditional Random Field): Probabilistic Named Entity Recognition Training Corpus:

Feature Engineering:

- Is the word capitalized?
- Is the word in a list of city names?
- What is the previous word?
- What is the previous slot/tag?
- Word shape features (e.g., "Xx", "XX")

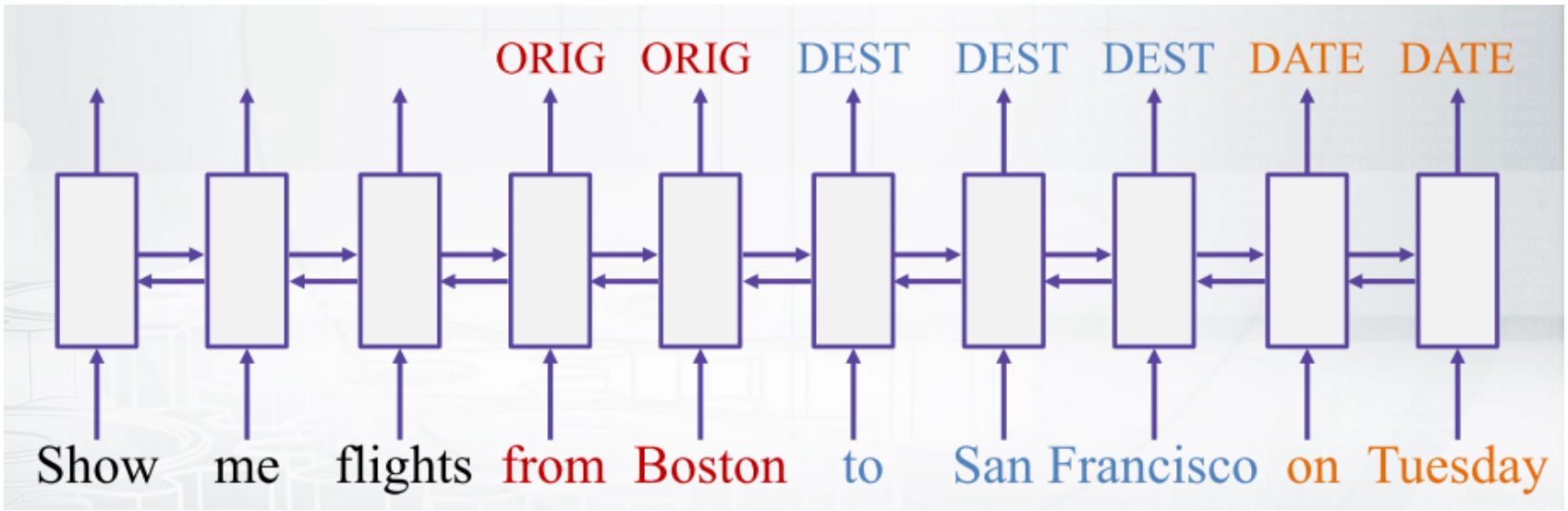
(Ref: https://www.coursera.org/learn/language-processing/)

YHK

Example: Deep Learning Approach

LSTM (Long Short-Term Memory):

- Large training corpus
- No manual feature engineering
- End-to-end architecture
- Learns representations automatically



(Ref: https://www.coursera.org/learn/language-processing/)

YHK

Example: Transformer Approach (Modern)

BERT/GPT for Slot Filling:

- Pre-trained on massive text corpora
- Fine-tuned on specific task
- Contextual embeddings for each token
- Attention mechanism captures dependencies
- State-of-the-art performance

Input: [CLS] Show me flights from Boston to San Francisco [SEP]
Output: O O O O B-ORIGIN O B-DEST I-DEST
Benefits: Transfer learning, minimal task-specific data needed

YHK

NLP Concepts

Core NLP Tasks: Text Processing

Document/Section Splitting

Correlation between three-dimensional ultrasound features and pathological prognostic factors in breast cancer

Sentence Splitting (Segmentation)

11 Sentences (= "T-" or "Terminable" units *only* if independent clauses are puctuated 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 #8201;= 0.014), a lower histological grade (P #8201;= 0.009) and positive ER or PR expression status (P #8201;= 0.001, 0.044).

Core NLP Tasks: Tokenization

Tokenization: Breaking text into tokens (words, subwords, characters)

- Word-level: "Hello world" $\rightarrow$ ["Hello", "world"]
- Challenges:
  - Hewlett-Packard (hyphen)
  - U.S.A. (periods)
  - Languages without spaces (Chinese, Japanese)
  - Contractions: don't $\rightarrow$ do + n't
- Modern approaches:
  - Subword tokenization (BPE, WordPiece)
  - SentencePiece for multilingual models
  - Handles out-of-vocabulary words better

YHK

Core NLP Tasks: Morphological Analysis

Stemming: Reduce words to stems

- detects, detected, detecting → detect
- Rules-based suffix stripping
- Porter's Algorithm (most popular for English)
- 36% reduction in vocabulary
- May produce non-words: sensitivities → sensit

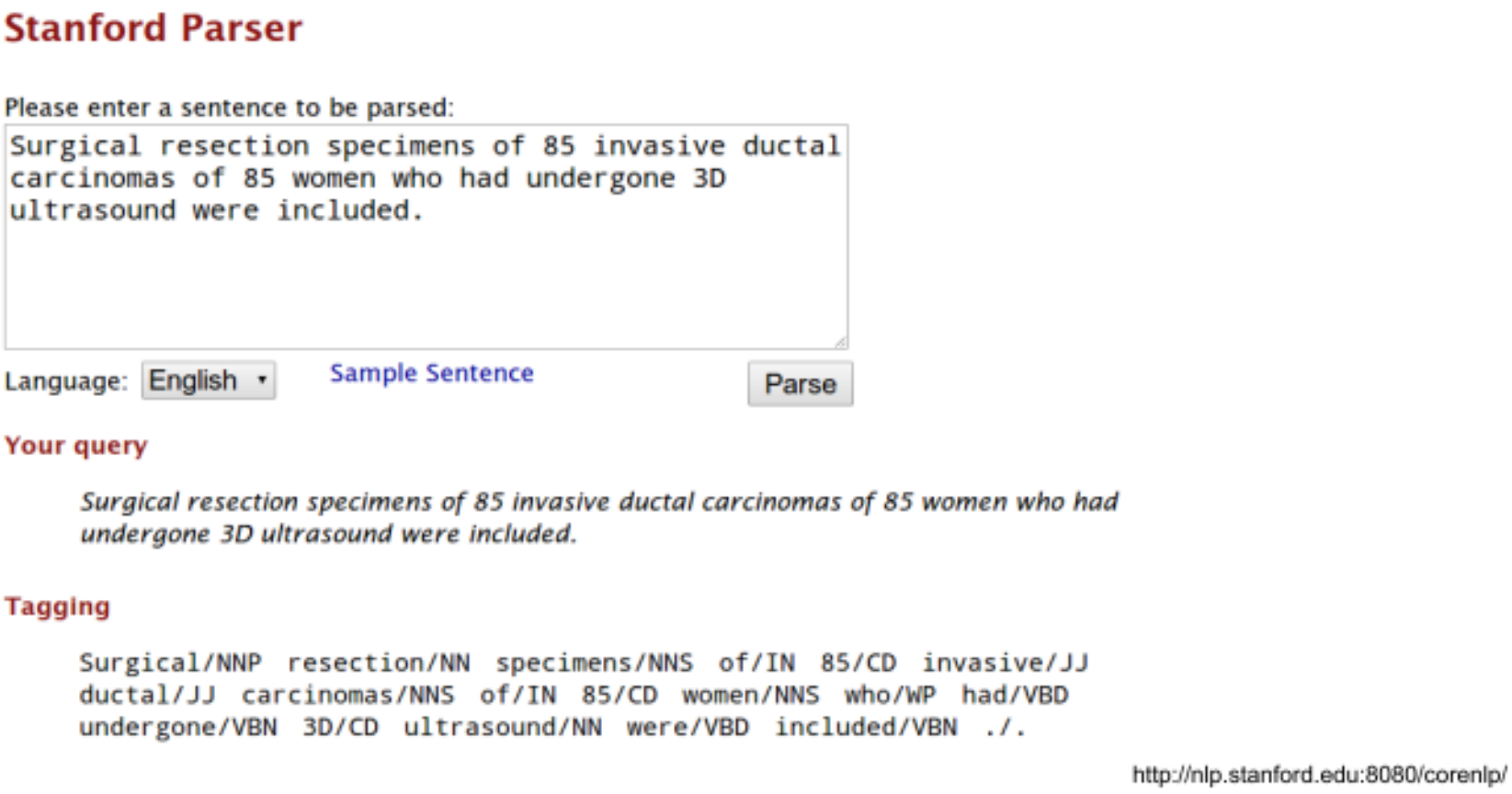
Lemmatization: Reduce to dictionary form

- Uses vocabulary and morphological analysis
- saw → see, been/was → be
- Linguistically correct lemmas
- Requires part-of-speech information

YHK

Core NLP Tasks: Part-of-Speech Tagging

POS Tagging: Assign syntactic tags to words



Tags include: Noun, Verb, Adjective, Adverb, Preposition, etc.
Applications:

- Prerequisite for parsing

YHK

Core NLP Tasks: Syntactic Parsing

Parsing: Build 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/

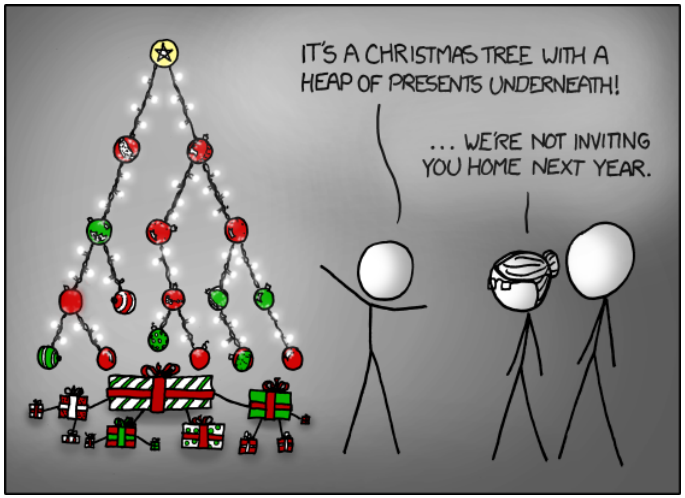
Types:

- Constituency parsing (phrase structure)
- Dependency parsing (word relationships)

YHK

Syntax Example

"DaimlerChrysler's shares rose three eighths to 22"



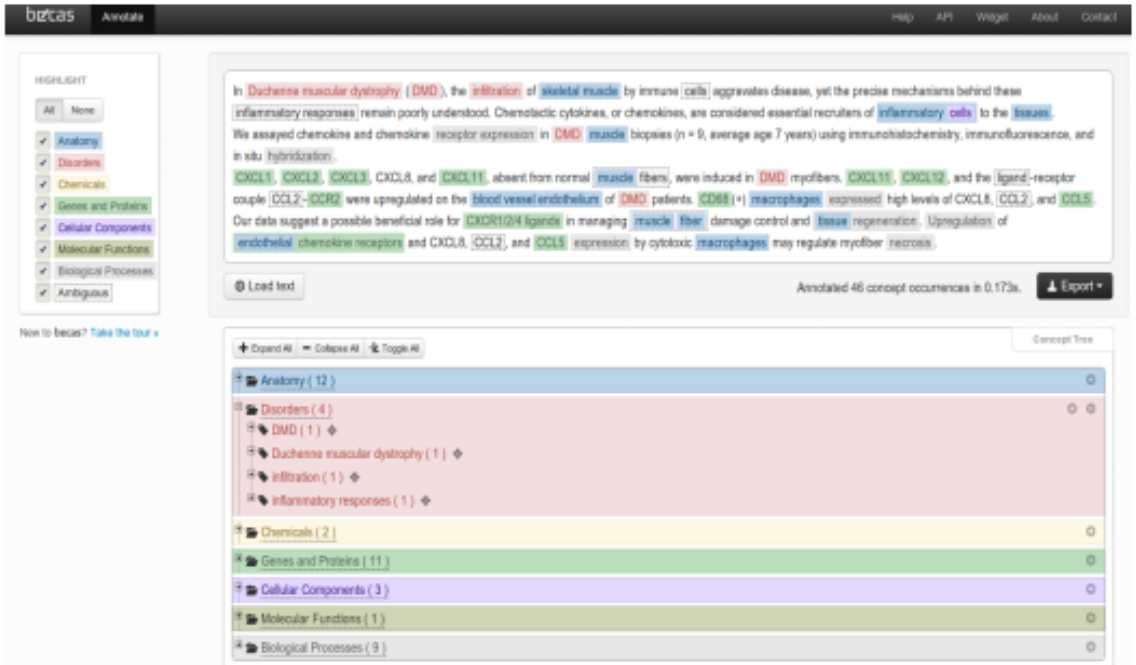
Sample grammar rules:

$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid TWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

YHK

Core NLP Tasks: Named Entity Recognition

NER: Identify and classify named entities



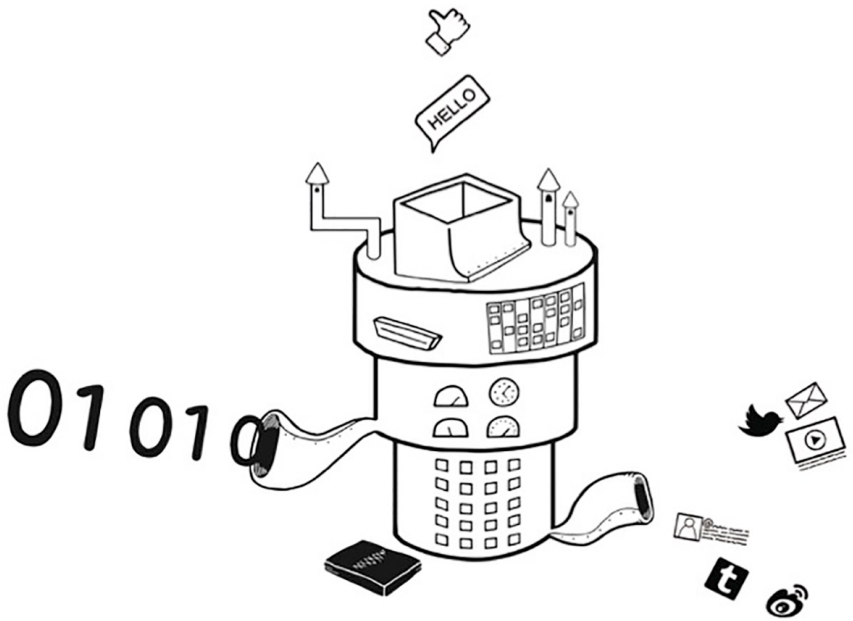
Common entity types:

- PERSON, ORGANIZATION, LOCATION
- DATE, TIME, MONEY
- PRODUCT, EVENT

Modern approaches: BERT-based models, few-shot learning with LLMs

YHK

Embedding



- Convert letters, words, and ideas into numbers
- Once we have numbers, we can use mathematics and machine learning
- Humans like words, computers like numbers

YHK

Word Representations: The Evolution

Traditional: One-Hot Encoding

- $Apple = [1, 0, 0, 0, ..., 0]$
- $Orange = [0, 1, 0, 0, ..., 0]$
- Problems: Sparse, no semantic meaning, huge dimensionality

Modern: Dense Embeddings (Word2Vec, GloVe)

- Each word $\rightarrow$ dense vector (e.g., 300 dimensions)
- Similar words have similar vectors
- $king - man + woman \approx queen$

Latest: Contextual Embeddings (BERT, GPT)

- Same word, different contexts $\rightarrow$ different vectors
- "bank" in "river bank" vs "savings bank"
- Captures polysemy and context

YHK

Word2Vec: Distributed Representations

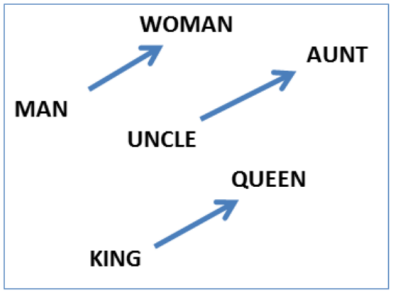
Key properties:

- Dense vectors (typically 100-300 dimensions)
- Trained on large corpora (unsupervised)
- Captures semantic relationships
- Two architectures: CBOW, Skip-gram

YHK

Word Embeddings: Semantic Relationships

Gender relation:



Country-capital relation:

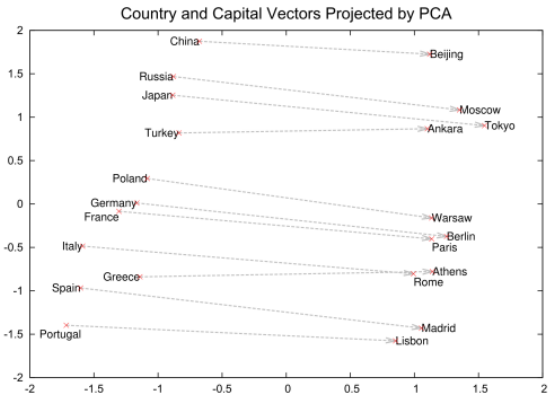


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

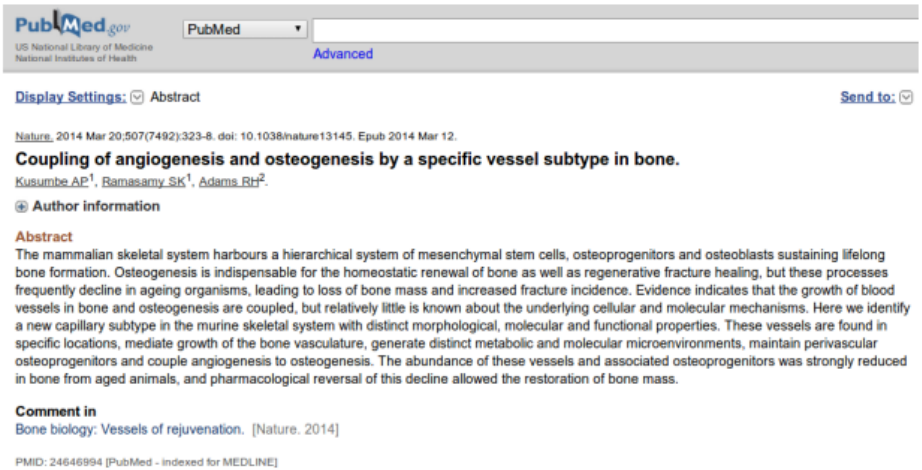
Vector arithmetic captures analogies!

YHK

NLP Applications

NLP Applications: Text Classification

Text Categorization: Assign predefined categories



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, osteoprogenitors and osteoblasts sustaining lifelong bone formation. Osteogenesis is indispensable for the homeostatic renewal of bone as well as regenerative fracture healing, but these processes frequently decline in ageing 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.

Comment in

Bone biology: Vessels of rejuvenation. [Nature. 2014]

PMID: 24646994 [PubMed - indexed for MEDLINE]

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

Applications:

- Spam detection
- Sentiment analysis
- Topic classification
- Intent detection

YHK

NLP Applications: Sentiment Analysis

Sentiment Analysis: Identify opinions and emotions

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

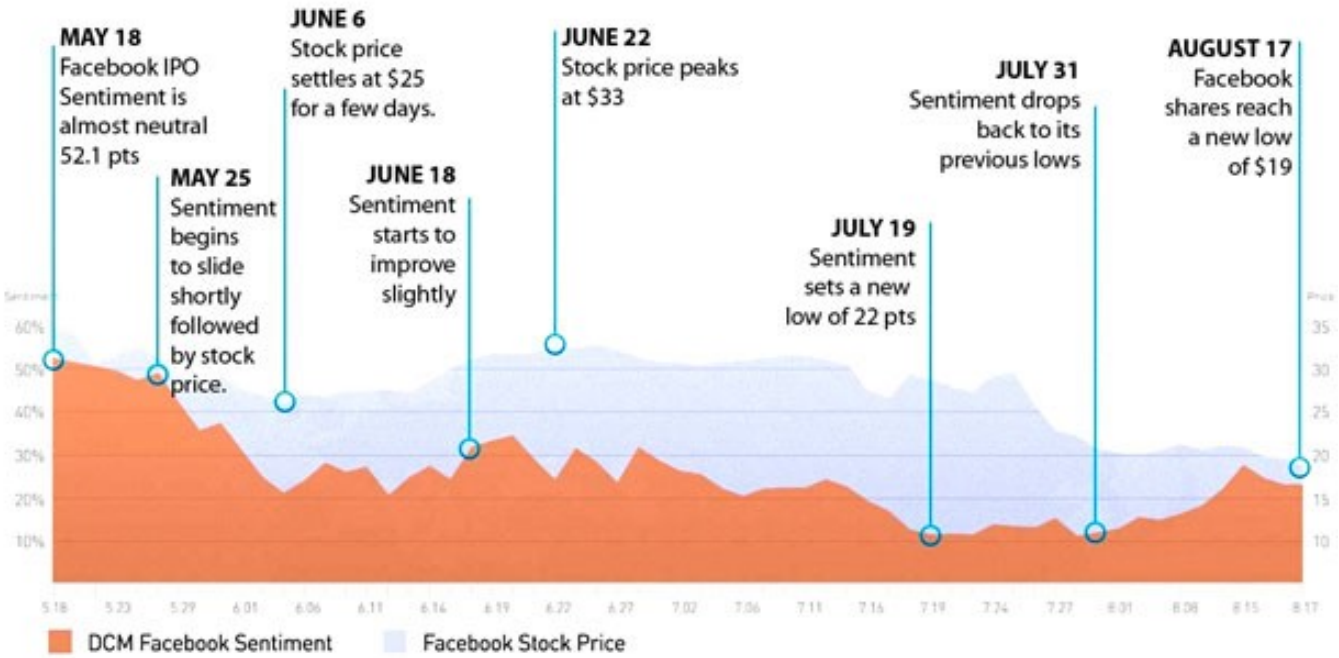
› See more 3 star, 2 star, 1 star reviews

Use cases:

- Product reviews
- Social media monitoring
- Customer feedback analysis
- Financial market sentiment

YHK

Sentiment Analysis in Finance



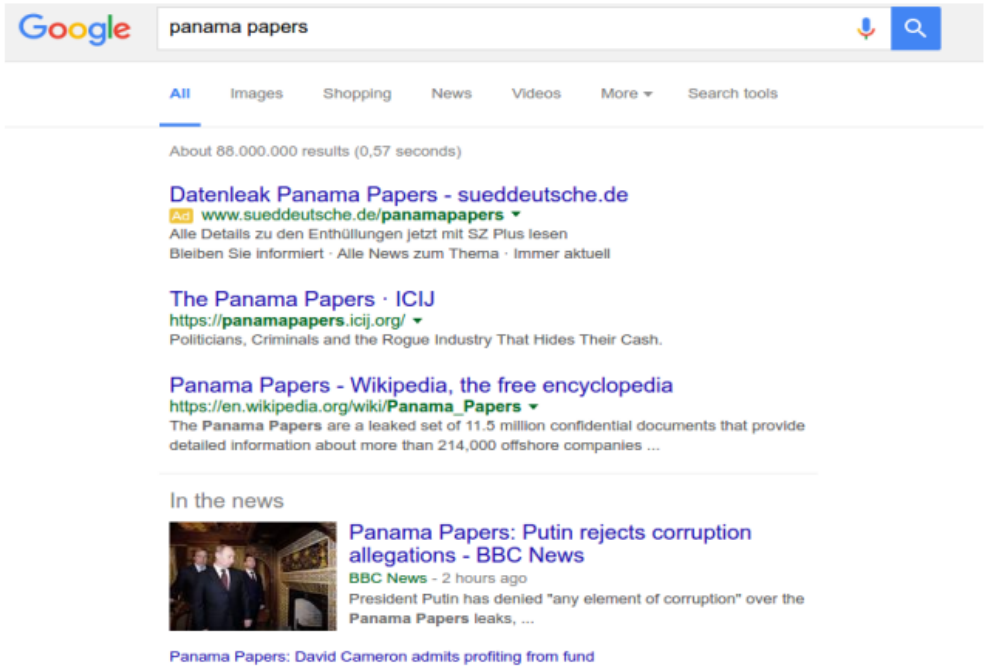
Applications:

- Stock market prediction
- News impact analysis
- Earnings call sentiment
- Social media financial sentiment

YHK

NLP Applications: Information Retrieval

Information Retrieval: Find relevant information



Components:

- Query understanding
- Document ranking
- Semantic search (beyond keyword matching)
- Modern: Dense retrieval with embeddings

YHK

NLP Applications: Information Extraction

Information Extraction: Extract structured information

Merkel at the EPP Summit, March 2016

Chancellor of Germany	
Incumbent	
Assumed office	22 November 2005
President	Horst Köhler Christian Wulff Joachim Gauck
Deputy	Franz Müntefering Frank-Walter Steinmeier Guido Westerwelle Philipp Rösler Sigmar Gabriel
Preceded by	Gerhard Schröder
Leader of the Christian Democratic Union	
Incumbent	
Assumed office	

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 Nolte
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	

Tasks:

- Named Entity Recognition
- Relation Extraction
- Event Extraction
- Template Filling

YHK

NLP Applications: Machine Translation

Machine Translation: Translate between languages

Evolution:

- Rule-based MT (1950s-1980s)
- Statistical MT (1990s-2010s)
- Neural MT (2014-present)
- Transformer-based (2017-present): Google Translate, DeepL

YHK

NLP Applications: Question Answering

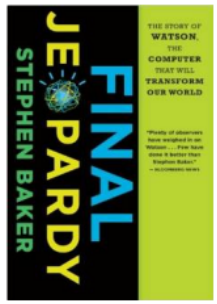
Question Answering: Provide direct answers to questions

===> 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.

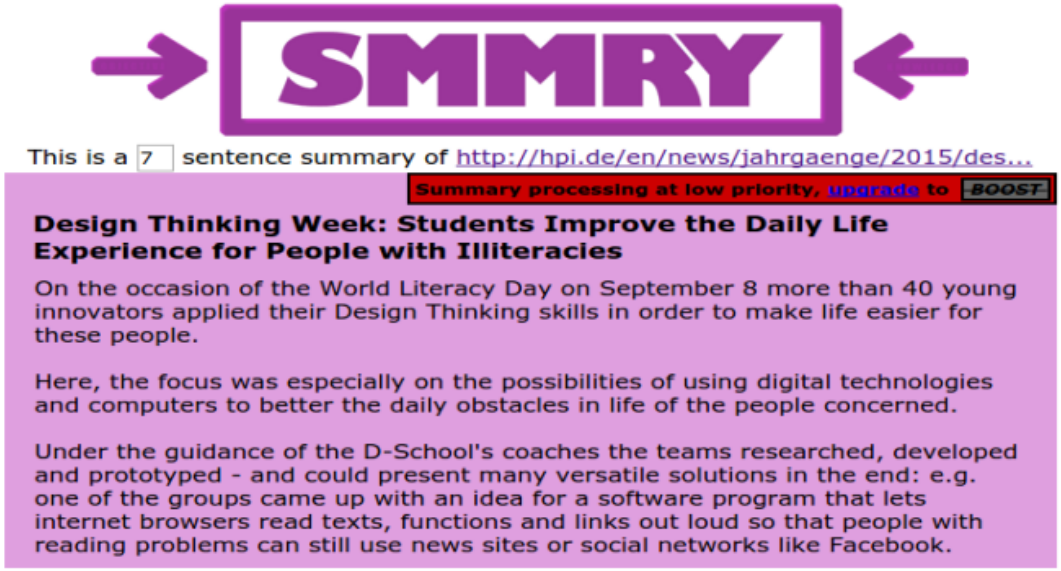
IBM Watson in Jeopardy (2011):



YHK

NLP Applications: Text Summarization

Summarization: Generate concise summaries



http://smmry.com/

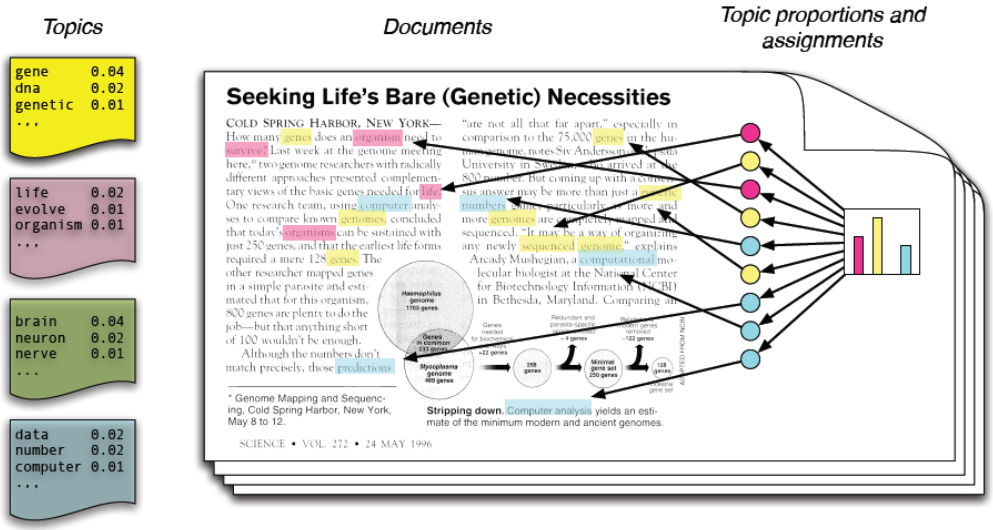
Types:

- Extractive: Select important sentences
- Abstractive: Generate new summary text
- Query-focused: Summarize w.r.t. specific query
- Modern: Transformer-based abstractive (BART, T5, GPT)

YHK

NLP Applications: Topic Modeling

Topic Modeling: Discover abstract topics in documents



Methods:

- Latent Dirichlet Allocation (LDA)
- Non-negative Matrix Factorization (NMF)
- BERTopic (modern approach)

YHK

NLP Applications: Modern Capabilities

Recent Breakthroughs with Large Language Models:

- Code generation (GitHub Copilot, GPT-4)
- Creative writing (stories, poems, scripts)
- Conversational AI (ChatGPT, Claude, Gemini)
- Multimodal understanding (text + images)
- Reasoning and problem-solving
- Few-shot and zero-shot learning

The field is rapidly evolving!

YHK

Modern NLP Stack (2025)

Foundation Models:

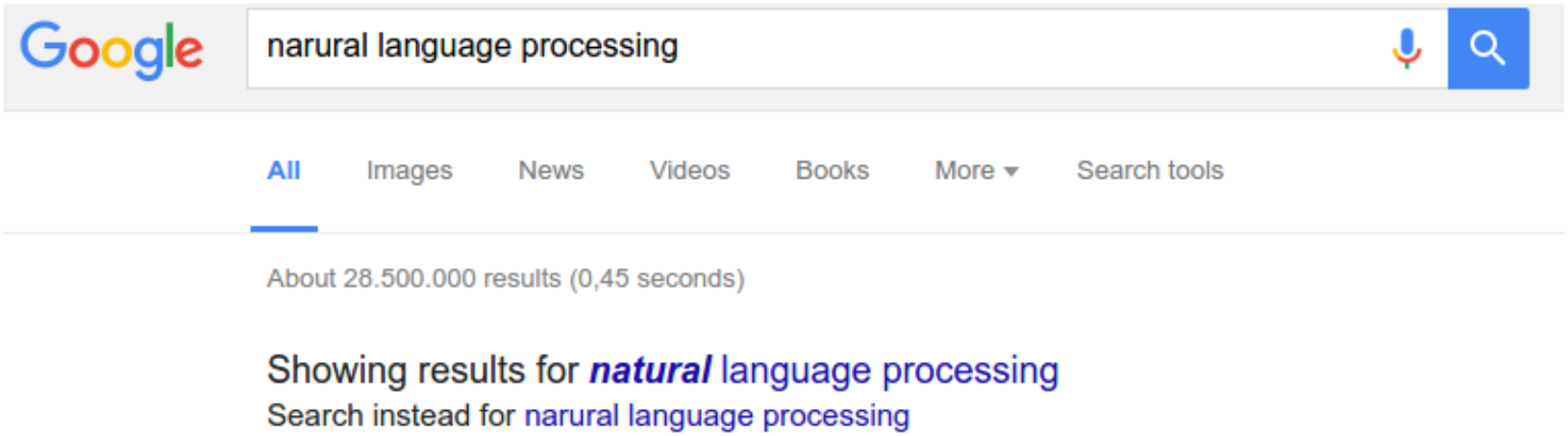
- GPT-4, GPT-4o (OpenAI)
- Claude 3.5 Sonnet, Claude 4 (Anthropic)
- Gemini 1.5, Gemini 2.0 (Google)
- LLaMA 3 (Meta)
- Mistral, Mixtral (Mistral AI)

Key Technologies:

- Transformers with billions/trillions of parameters
- Retrieval-Augmented Generation (RAG)
- Reinforcement Learning from Human Feedback (RLHF)
- Constitutional AI and alignment techniques
- Multimodal architectures

YHK

Spell and Grammar Checking



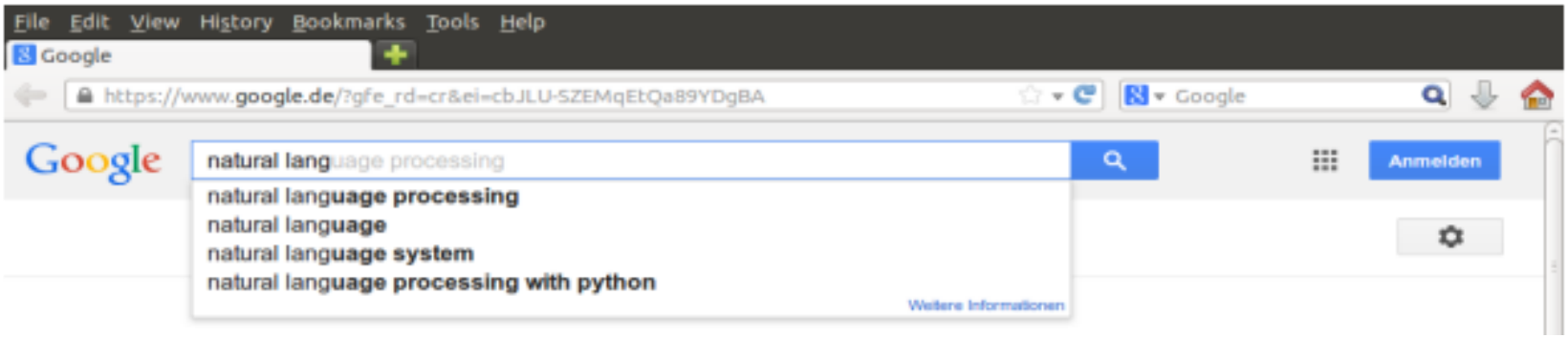
Tasks:

- Spelling error detection and correction
- Grammar error detection
- Style suggestions
- Modern: Context-aware corrections with LLMs

YHK

Word Prediction and Autocomplete

Word Prediction: Predict next likely word

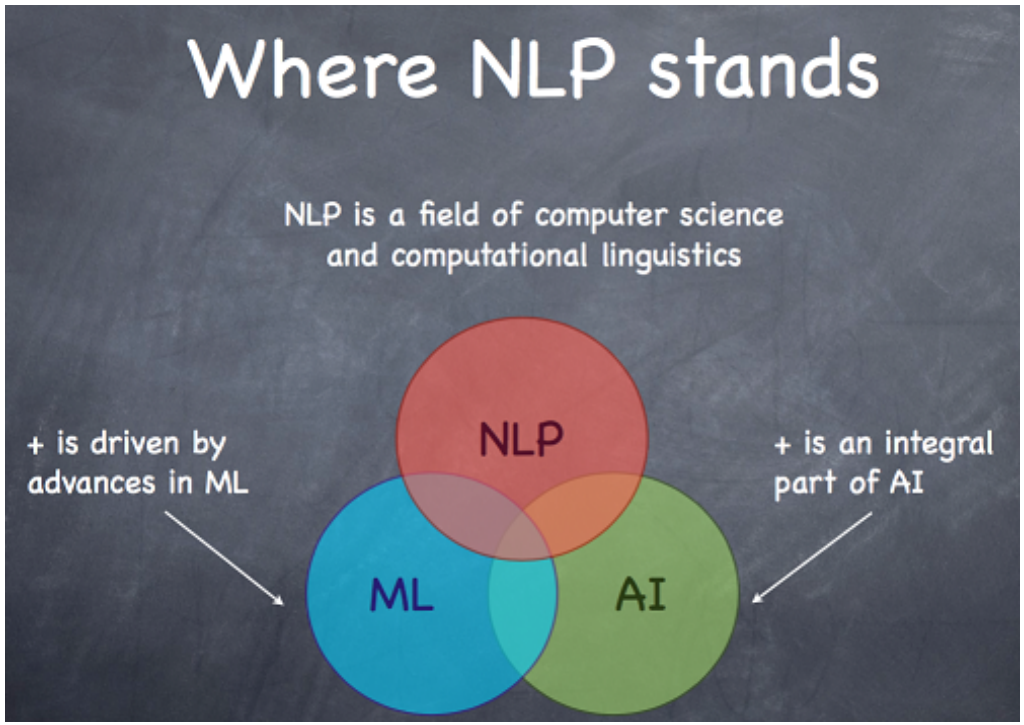


Applications:

- Mobile keyboard typing
- Search engine suggestions
- Email composition
- Code completion

YHK

Current State of NLP



NLP is embedded everywhere:

- Search engines
- Virtual assistants
- Customer service
- Content moderation
- Healthcare (clinical NLP)
- Legal (contract analysis)

YHK

Level of Difficulty: Task Classification

Mostly Solved:

- Spell checking and grammar correction
- Part-of-speech tagging
- Named entity recognition (common entities)
- Basic text classification

Good Progress:

- Machine translation
- Sentiment analysis
- Information retrieval
- Information extraction
- Text summarization

Still Challenging:

- Complex reasoning and inference
- Common sense understanding
- Long-form coherent generation
- Multilingual low-resource languages
- Factual consistency and hallucination control

YHK

NLP Trends in 2025

Trends driving NLP adoption:

- Enormous amount of machine-readable text
  - Newspapers, web pages, social media
  - Medical records, financial filings
  - Product reviews, discussion forums
- Conversational agents as primary interface
  - ChatGPT, Claude, virtual assistants
- Human-human interaction mediated by computers
  - Social media, messaging platforms

This means copious data available for NLP system development!

YHK

The Promise of NLP

Growing importance in multiple arenas:

- Scientific: accelerating research
- Economic: business intelligence, automation
- Social: breaking language barriers
- Cultural: preserving and translating heritage

Wide range of stakeholders need NLP knowledge:

- Academia: linguistics, computer science, AI
- Industry: HCI, business analysis, web development
- Healthcare: clinical documentation, diagnosis support
- Legal: contract analysis, legal research

Goal: Open NLP to a broad audience

YHK

NLP and Intelligence

- Long-standing challenge: build intelligent machines
- Chief measure of machine intelligence: linguistic capability
- Turing Test: conversational ability as intelligence criterion

Example human-machine dialog:

S: How may I help you?
U: When is Saving Private Ryan playing?
S: For what theater?
U: The Paramount theater.
S: Saving Private Ryan is not playing at the Paramount theater, but it's playing at the Madison theater at 3:00, 5:30, and 10:30.

YHK

Current Limitations

Today's systems limited to narrow domains:

- Can't easily add new capabilities
- Require domain-specific training
- Common-sense reasoning still challenging
- Hallucination and factual errors
- Lack of true understanding vs pattern matching

To improve, we need:

- Better knowledge representation
- Improved reasoning capabilities
- More robust grounding in reality
- Continued research in alignment and safety

YHK

Conclusion & Outlook

Key Takeaways

- NLP is about making computers understand and generate human language
- Language is inherently ambiguous and context-dependent
- Multiple approaches: rules, ML, deep learning, LLMs
- Huge progress in recent years, but challenges remain
- Applications everywhere: search, translation, assistants, analysis

In Nutshell

- NLP is an effort to do useful things with natural language
- Humans like words, computers like numbers
- We bridge this gap through representation learning
- Modern NLP combines:
  - Large data
  - Powerful compute
  - Advanced algorithms (Transformers, LLMs)
- Exciting time to be in NLP - rapid innovation!

YHK

Resources and Next Steps

Learn More:

- Stanford CS224N: Natural Language Processing with Deep Learning
- fast.ai NLP course
- Hugging Face tutorials and documentation
- Papers: arxiv.org/list/cs.CL (Computation and Language)

Practice:

- Kaggle NLP competitions
- Build projects with Hugging Face Transformers
- Experiment with OpenAI API, Anthropic Claude
- Contribute to open-source NLP libraries

YHK

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

YHK

Thanks . . .

- Office Hours: Saturdays, 3 to 5 pm (IST); Free-Open to all; email for appointment to yogeshkulkarni at yahoo dot com
- Call + 9 1 9 8 9 0 2 5 1 4 0 6

(https://www.linkedin.com/in/yogeshkulkarni/)

(https://medium.com/@yogeshharibhaukulkarni)

(https://www.github.com/yogeshhk/)

YHK