Introduction to Natural Language Processing & Computational Linguistics

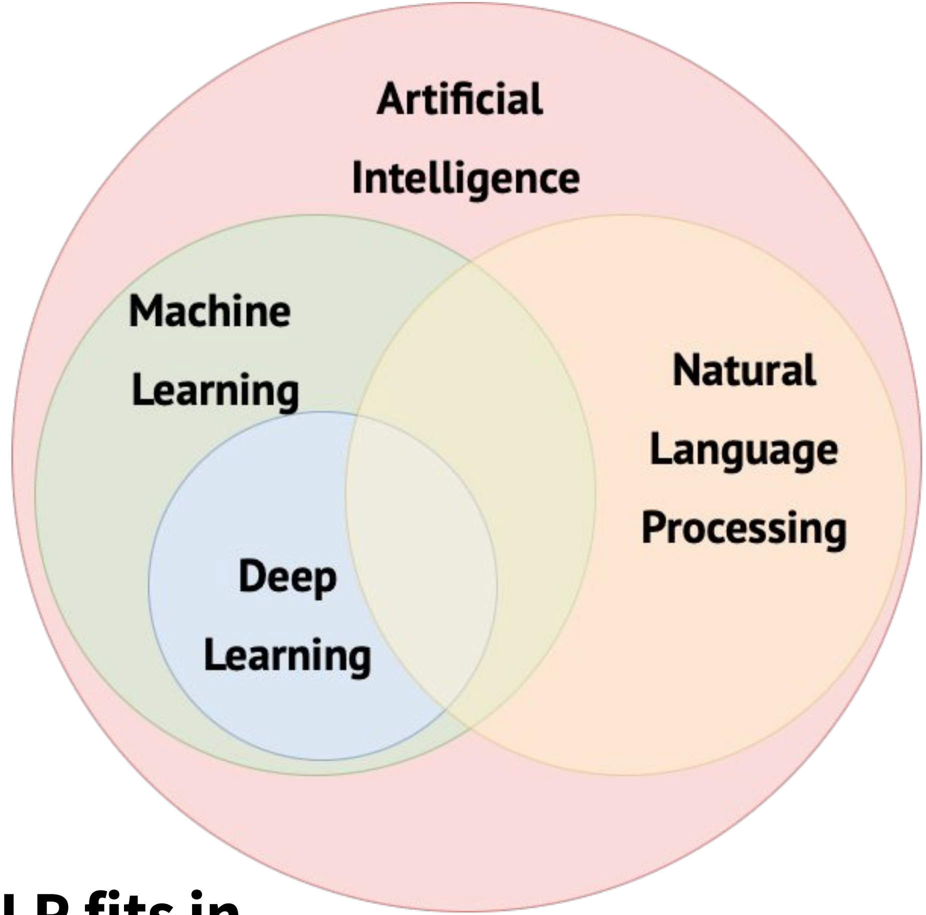
Lecture 1: 21-479-0105 Computational Linguistics

Dr. Jeena Kleenankandy

07 August 2023

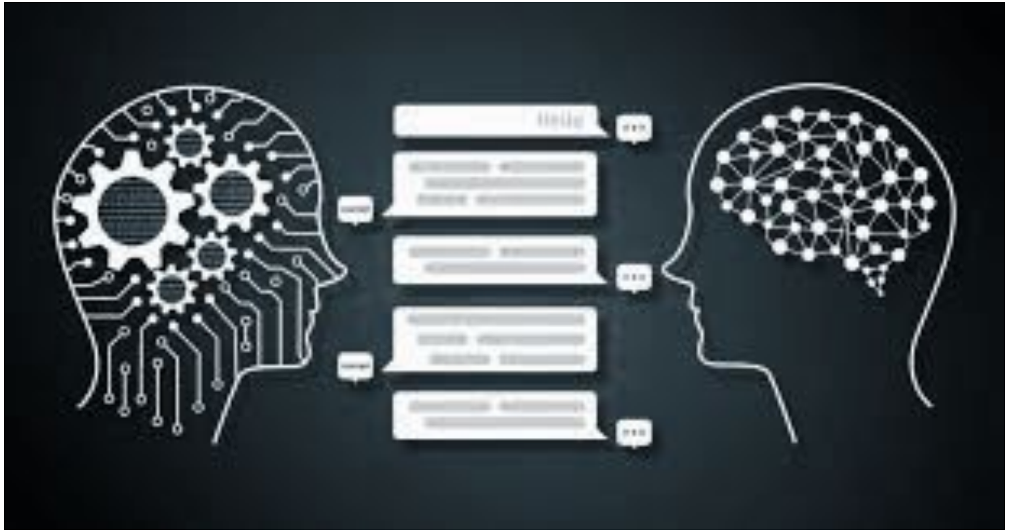
Agenda

- Introduction to NLP
- Key NLP tasks
- Application of NLP in various domains
  - Healthcare
  - E-commerce
  - Social media
  - Human Resource Management
  - Legal Domain
  - Finance/Banking
- Why NLP is challenging



Where NLP fits in..

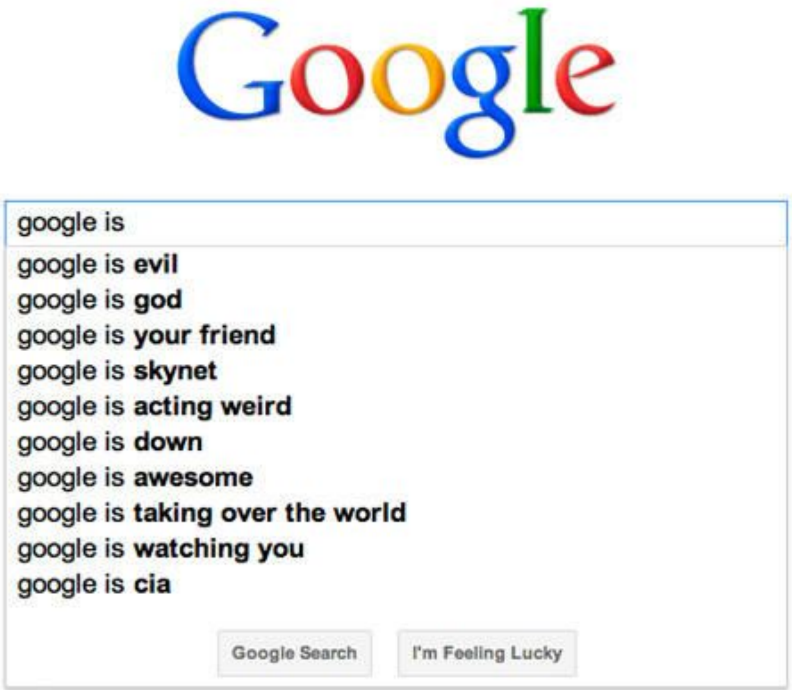
What is Natural Language Processing ?



- Designing/building computational models to *understand* and *generate* human languages to get some useful task done

NLP in our daily life...

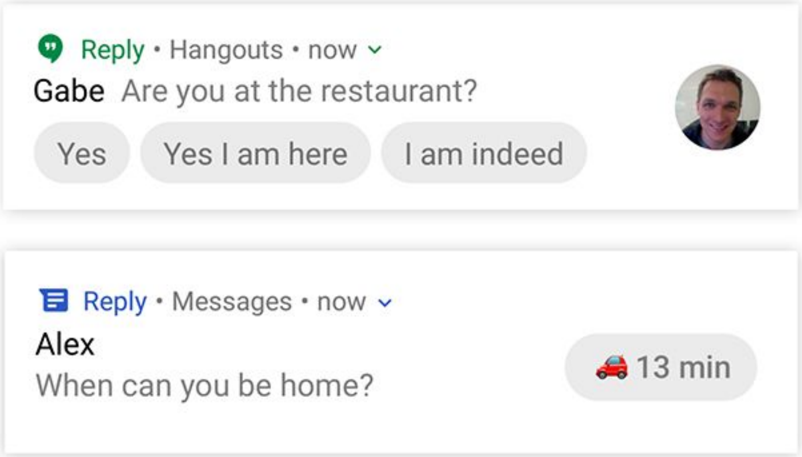
Google Search
Google Autocomplete
Personalised Ads
Video suggestions
Chat bots



google is

google is evil
google is god
google is your friend
google is skynet
google is acting weird
google is down
google is awesome
google is taking over the world
google is watching you
google is cia

NLP in our daily life...

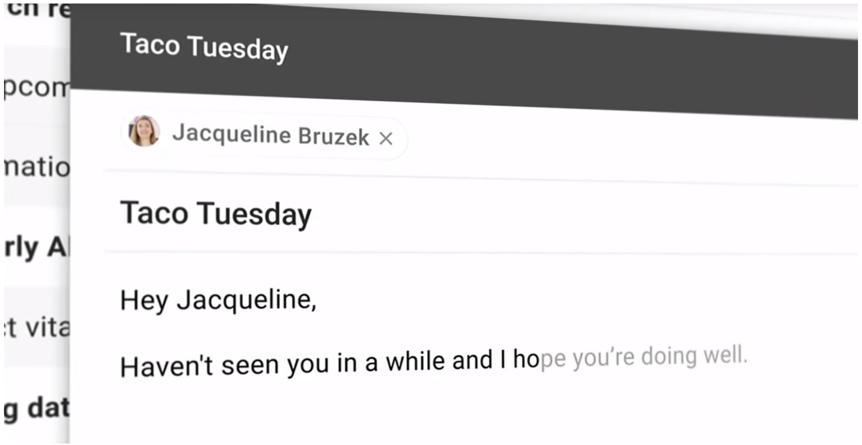


Gabe Are you at the restaurant?

Alex
When can you be home?

Google's Smart reply

Gmail's Smart Compose



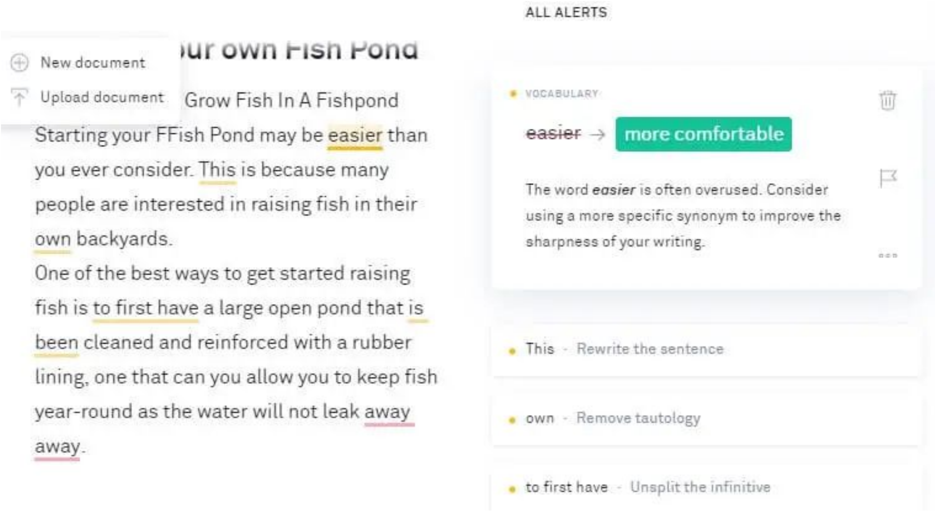
Taco Tuesday

Hey Jacqueline,

Haven't seen you in a while and I hope you're doing well.

Spam filter

NLP in our daily life...



ur own Fish Pond

Grow Fish In A Fishpond

Starting your FFish Pond may be easier than you ever consider. This is because many people are interested in raising fish in their own backyards.

One of the best ways to get started raising fish is to first have a large open pond that is been cleaned and reinforced with a rubber lining, one that can you allow you to keep fish year-round as the water will not leak away away.

ALL ALERTS

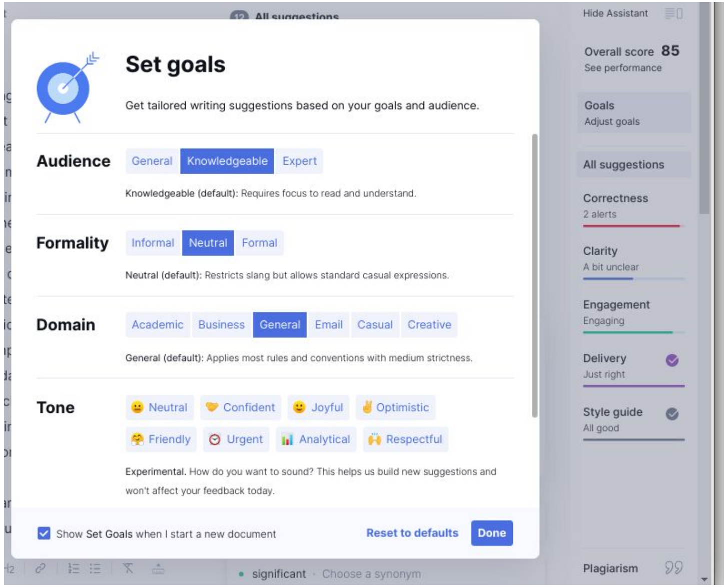
VOCABULARY

easier → more comfortable

The word *easier* is often overused. Consider using a more specific synonym to improve the sharpness of your writing.

- This · Rewrite the sentence
- own · Remove tautology
- to first have · Unsplit the infinitive

Grammarly - writing assistant



Set goals

Get tailored writing suggestions based on your goals and audience.

Audience: General [✓ Knowledgeable] Expert

Knowledgeable (default): Requires focus to read and understand.

Formality: Informal [✓ Neutral] Formal

Neutral (default): Restricts slang but allows standard casual expressions.

Domain: Academic Business [✓ General] Email Casual Creative

General (default): Applies most rules and conventions with medium strictness.

Tone: Neutral Confident Joyful Optimistic Friendly Urgent Analytical Respectful

Experimental. How do you want to sound? This helps us build new suggestions and won't affect your feedback today.

☑ Show Set Goals when I start a new document

Overall score 85
See performance

Goals
Adjust goals

All suggestions

Correctness
2 alerts

Clarity
A bit unclear

Engagement
Engaging

Delivery
Just right

Style guide
All good

Plagiarism

significant · Choose a synonym

Not just Grammar correction!

NLP in our daily life...

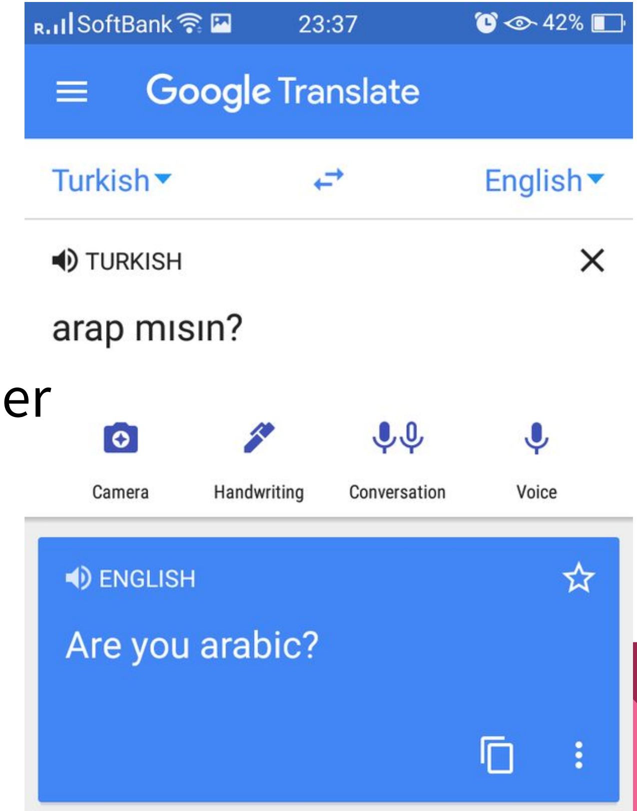
Personal assistants using voice commands

Speech Recognition

NLP in our daily life...

Language Translation

Converts text from one language to another



R.ıl SoftBank 23:37 42%

Google Translate

Turkish English

TURKISH

arap mısın?

ENGLISH

Are you arabic?

Key NLP tasks

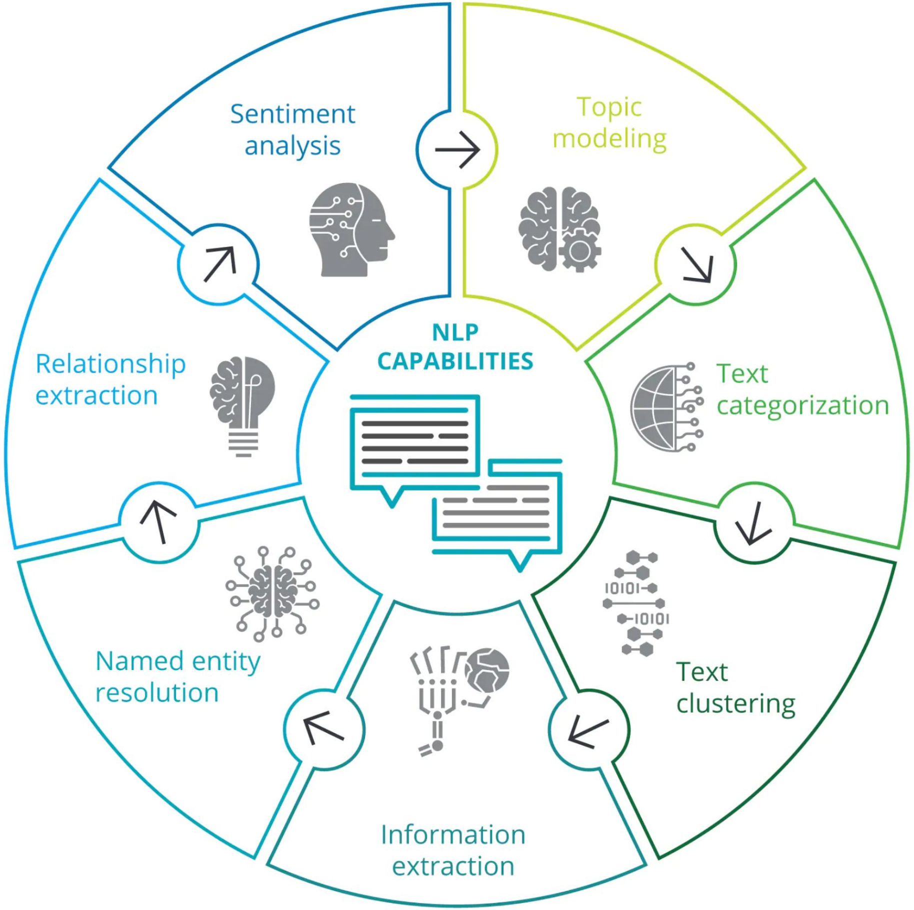


Image courtesy:
https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html

NLP Applications

NLP in Healthcare

Named Entity Recognition (NER) and Relation Extraction (RE) in Electronic Health Records (EHR) :

Disease Risk Prediction

- facilitate earlier detection of diseases and potentially improving patient outcomes.

Personalized Healthcare

- suggesting personalized treatment plans
- Chatbot therapist helping people with anxiety and other disorders.

Disease Evolution Prediction

Drug Reaction Detection

Extracting social determinants of health issues

NLP in ecommerce & digital marketing

Sentiment Analysis, Machine Translation and Language understanding in:

- Understanding User Intent
- Semantics for Search Engine Experience
- Autocorrect and Autocomplete
- Virtual Assistants & Chat-based Product Recommendations
- Target Advertising
- Customer Review Analysis
- Translation for Global Reach

NLP and Social media content

Identifying fake news and hate speech in posts involves

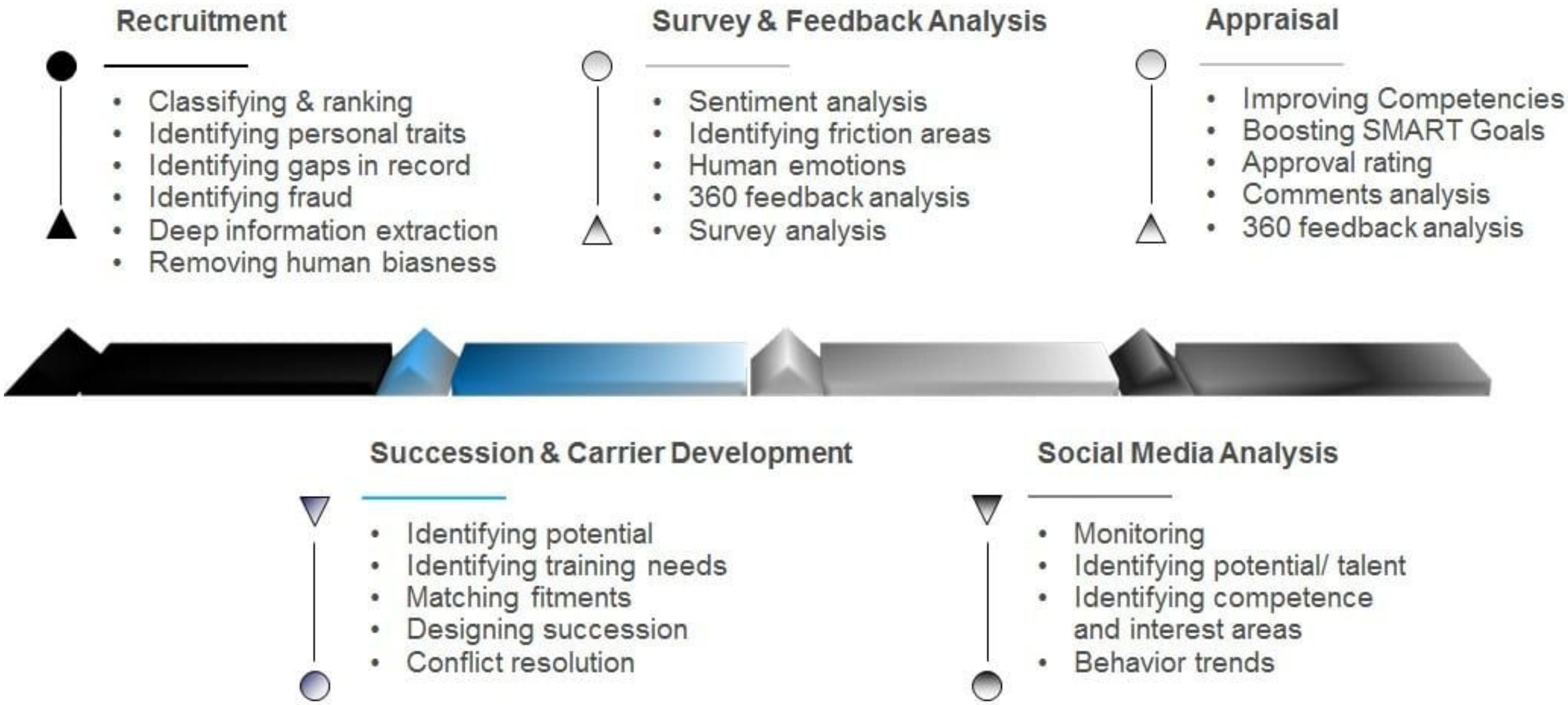
- Stance detection
- Automatic summarization
- Fact checking
- Sentiment Analysis

Stock market prediction

- Tracking news, reports, comments about possible mergers between companies
- Sentiment Analysis on Twitter & other social media platforms

Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective, Qi Su, Mingyu Wan Xiaoqian Liu, Chu-Ren Huang

NLP and Human Resource Management



Recruitment

- Classifying & ranking
- Identifying personal traits
- Identifying gaps in record
- Identifying fraud
- Deep information extraction
- Removing human biasness

Survey & Feedback Analysis

- Sentiment analysis
- Identifying friction areas
- Human emotions
- 360 feedback analysis
- Survey analysis

Appraisal

- Improving Competencies
- Boosting SMART Goals
- Approval rating
- Comments analysis
- 360 feedback analysis

Succession & Carrier Development

- Identifying potential
- Identifying training needs
- Matching fitments
- Designing succession
- Conflict resolution

Social Media Analysis

- Monitoring
- Identifying potential/ talent
- Identifying competence and interest areas
- Behavior trends

Image courtesy :https://www.aihr.com/blog/natural-language-processing-revolutionize-human-resources/

NLP in Legal domain

Legal Judgment Prediction

Similar Case Matching

Legal Question Answering

Legal Document Summarization

Reviewing Legal contracts

Source: How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence
Haoxi Zhong,, Chaojun Xiao, , Cunchao Tu,, Tianyang Zhang, Zhiyuan Liu1, Maosong Sun

NLP in Finance

Credit Scoring

- assess creditworthiness of borrowers digital footprints across social media, browsing history, geo-location
- Fraud detection
- Claim Approval in Insurance

Customer Service

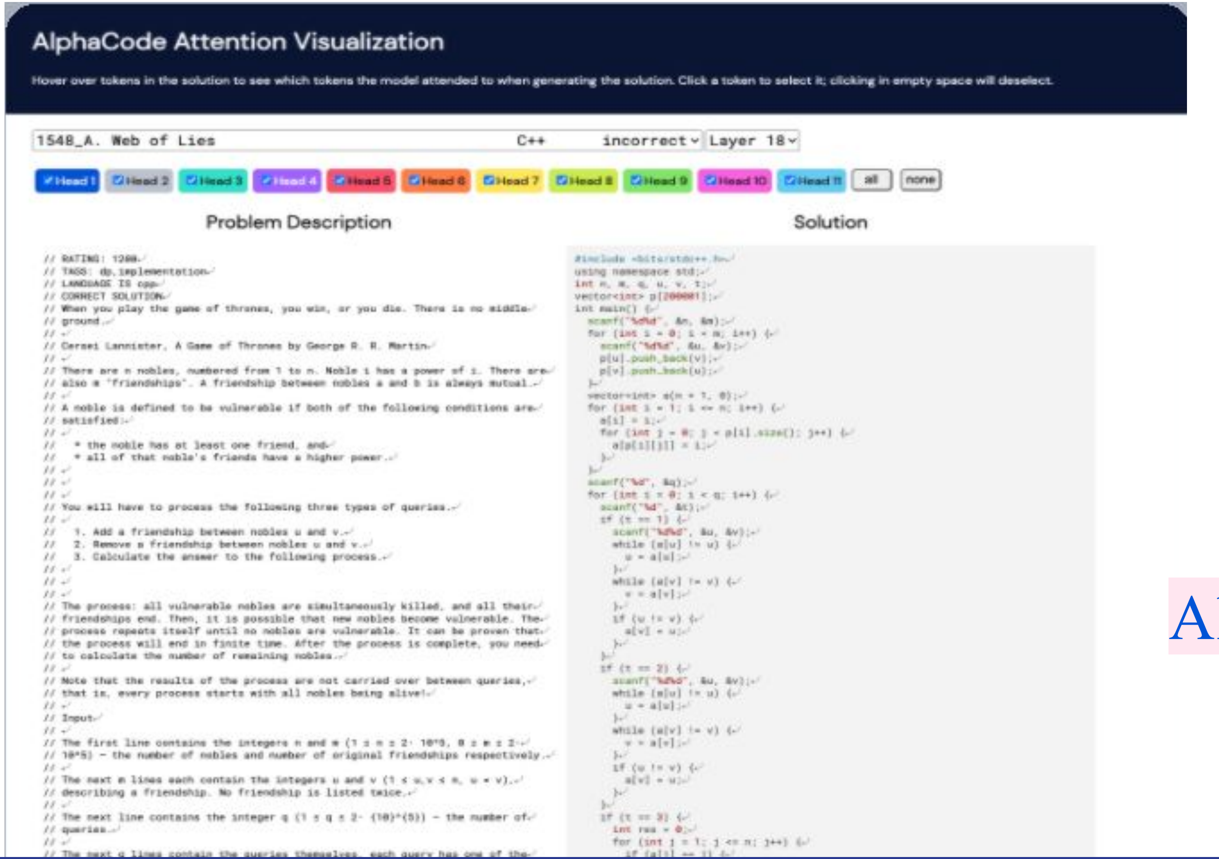
- Chatbots / virtual assistance
- savings $7.3 billion globally in operational cost in banking

Market Data Collection

- Stock market prediction

Analyse Banking contracts

NLP in Competitive programming



AlphaCode Attention Visualization

Hover over tokens in the solution to see which tokens the model attended to when generating the solution. Click a token to select it; clicking in empty space will deselect.

1548_A. Web of Lies | C++ | incorrect ▾ | Layer 18 ▾

Problem Description

```
// RATING: 1200↵
// TAGS: dp,implementation↵
// LANGUAGE IS cpp↵
// CORRECT SOLUTION↵
// When you play the game of thrones, you win, or you die. There is no middle↵
// ground.↵
// ↵
// Cersei Lannister, A Game of Thrones by George R. R. Martin↵
// ↵
// There are n nobles, numbered from 1 to n. Noble i has a power of i. There are↵
// also m "friendships". A friendship between nobles a and b is always mutual.↵
// ↵
// A noble is defined to be vulnerable if both of the following conditions are↵
// satisfied:↵
// ↵
//   * the noble has at least one friend, and↵
//   * all of that noble's friends have a higher power.↵
// ↵
// ↵
// ↵
// You will have to process the following three types of queries.↵
// ↵
//   1. Add a friendship between nobles u and v.↵
//   2. Remove a friendship between nobles u and v.↵
//   3. Calculate the answer to the following process.↵
// ↵
// ↵
// ↵
// The process: all vulnerable nobles are simultaneously killed, and all their↵
// friendships end. Then, it is possible that new nobles become vulnerable. The↵
// process repeats itself until no nobles are vulnerable. It can be proven that↵
// the process will end in finite time. After the process is complete, you need↵
// to calculate the number of remaining nobles.↵
// ↵
// Note that the results of the process are not carried over between queries,↵
// that is, every process starts with all nobles being alive!↵
// ↵
// Input↵
// ↵
// The first line contains the integers n and m (1 ≤ n ≤ 2· 10^5, 0 ≤ m ≤ 2·↵
// 10^5) — the number of nobles and number of original friendships respectively.↵
// ↵
// The next m lines each contain the integers u and v (1 ≤ u,v ≤ n, u ≠ v),↵
// describing a friendship. No friendship is listed twice.↵
// ↵
// The next line contains the integer q (1 ≤ q ≤ 2· (10)^(5)) — the number of↵
// queries.↵
// ↵
// The next q lines contain the queries themselves, each query has one of the↵
```

Solution

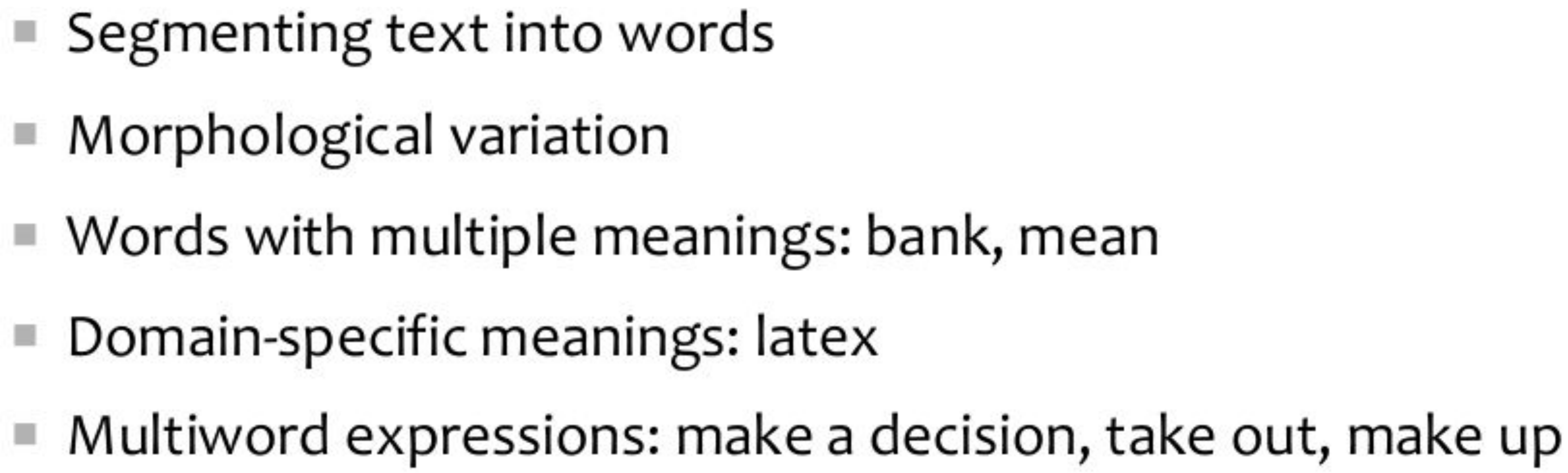
```
#include <bits/stdc++.h>↵
using namespace std;↵
int n, m, q, u, v, t;↵
vector<int> p[200001];↵
int main() {↵
  scanf("%d%d", &n, &m);↵
  for (int i = 0; i < m; i++) {↵
    scanf("%d%d", &u, &v);↵
    p[u].push_back(v);↵
    p[v].push_back(u);↵
  }↵
  vector<int> a(n + 1, 0);↵
  for (int i = 1; i <= n; i++) {↵
    a[i] = i;↵
    for (int j = 0; j < p[i].size(); j++) {↵
      a[p[i][j]] = i;↵
    }↵
  }↵
  scanf("%d", &q);↵
  for (int i = 0; i < q; i++) {↵
    scanf("%d", &t);↵
    if (t == 1) {↵
      scanf("%d%d", &u, &v);↵
      while (a[u] != u) {↵
        u = a[u];↵
      }↵
      while (a[v] != v) {↵
        v = a[v];↵
      }↵
      if (u != v) {↵
        a[v] = u;↵
      }↵
    }↵
    if (t == 2) {↵
      scanf("%d%d", &u, &v);↵
      while (a[u] != u) {↵
        u = a[u];↵
      }↵
      while (a[v] != v) {↵
        v = a[v];↵
      }↵
      if (u != v) {↵
        a[v] = u;↵
      }↵
    }↵
    if (t == 3) {↵
      int res = 0;↵
      for (int j = 1; j <= n; j++) {↵
        if (a[j] == j) {↵
```

AlphaCode

Why is NLP challenging ?

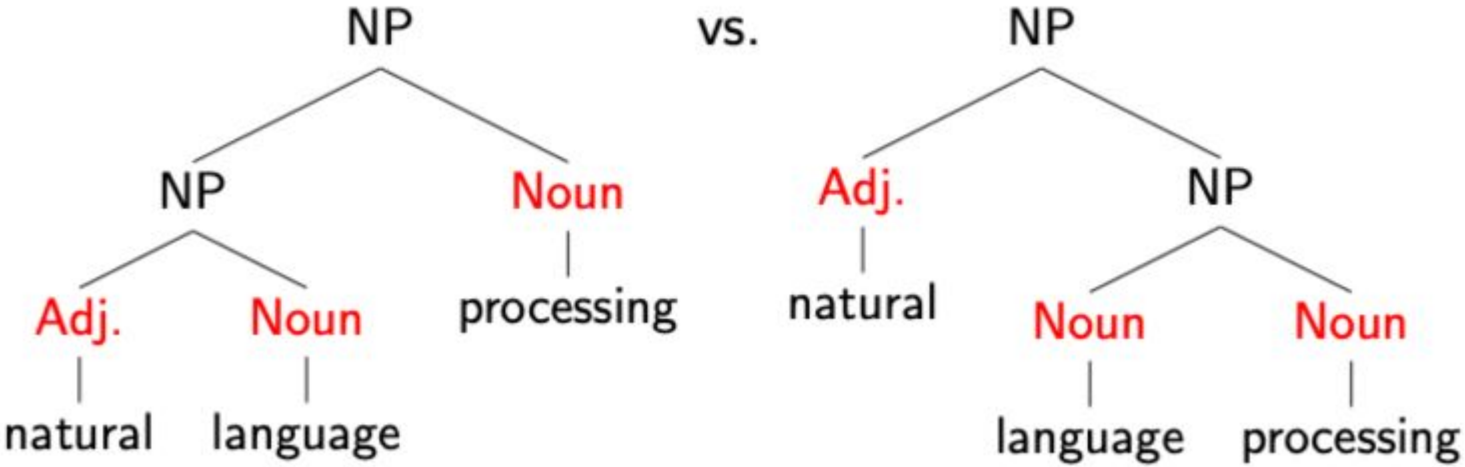
- Language is Ambiguous
    - Lexical, Syntactic, Semantic, Pragmatic etc.
- Language has Irregularities like
    - Sarcasm, idioms, metaphors and lot more..
- Language keeps evolving with variations across domains
- Linguistic communication is compressed
- Expressivity, unmodelled variables, unknown representations....

Even “words” are challenging

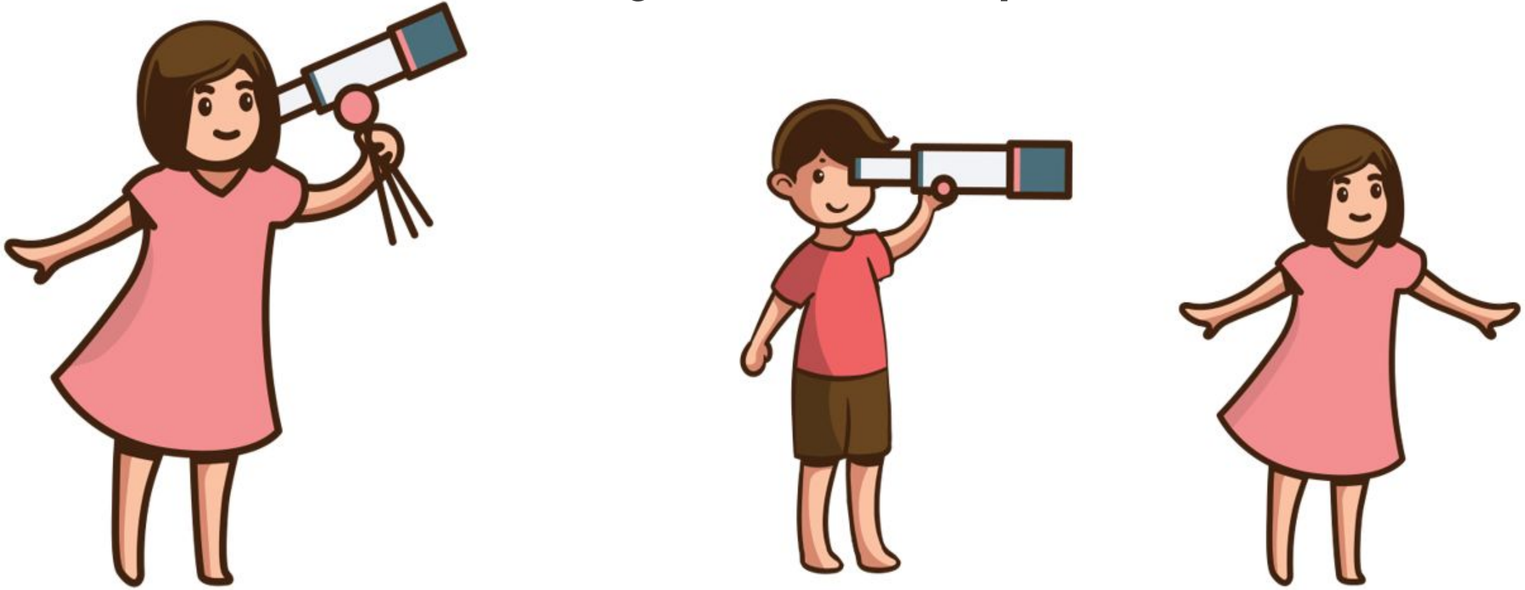


- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up

Syntactic ambiguity

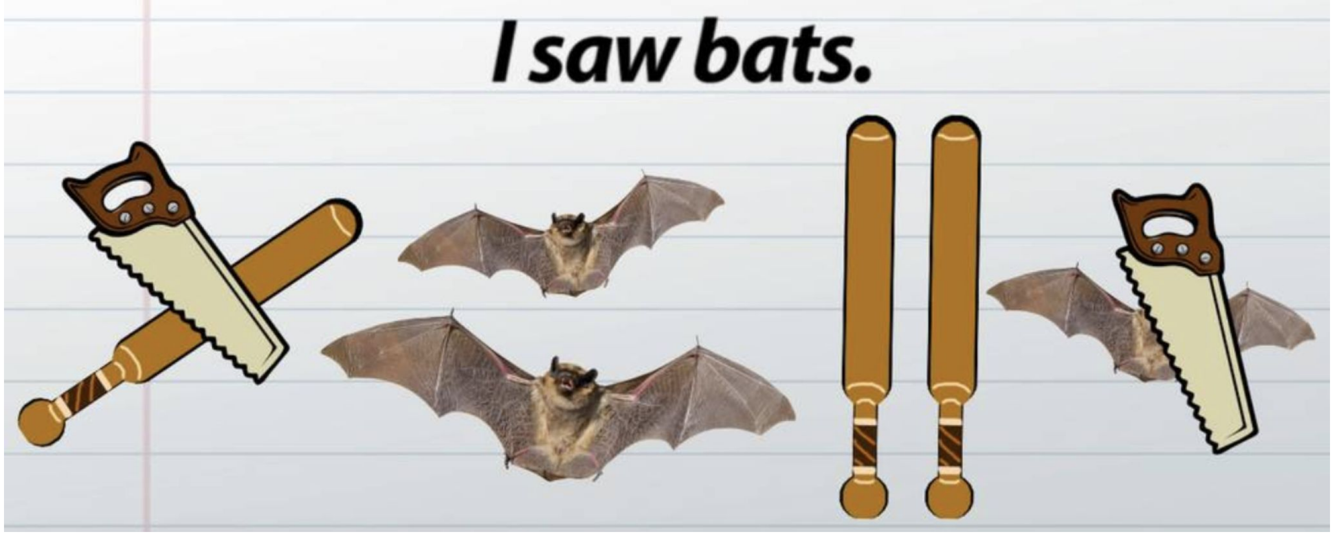


“I saw a girl with a telescope”



PP-attachment problem : An example of **Syntactic Ambiguity**

Semantic Ambiguity



Four different interpretation for a simple sentence !!

Irregularities…..

➔ He has a heart of gold

➔ Meet me at the bank.

➔ Kill two birds with one stone.

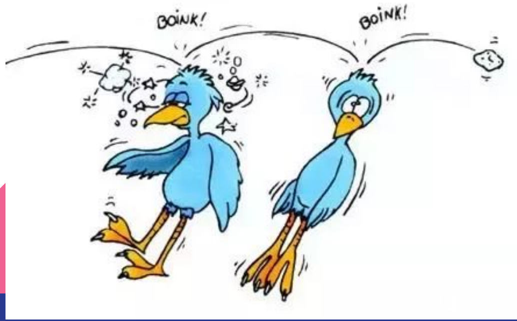
➔ It's a piece of cake

➔ It's raining cats and dogs

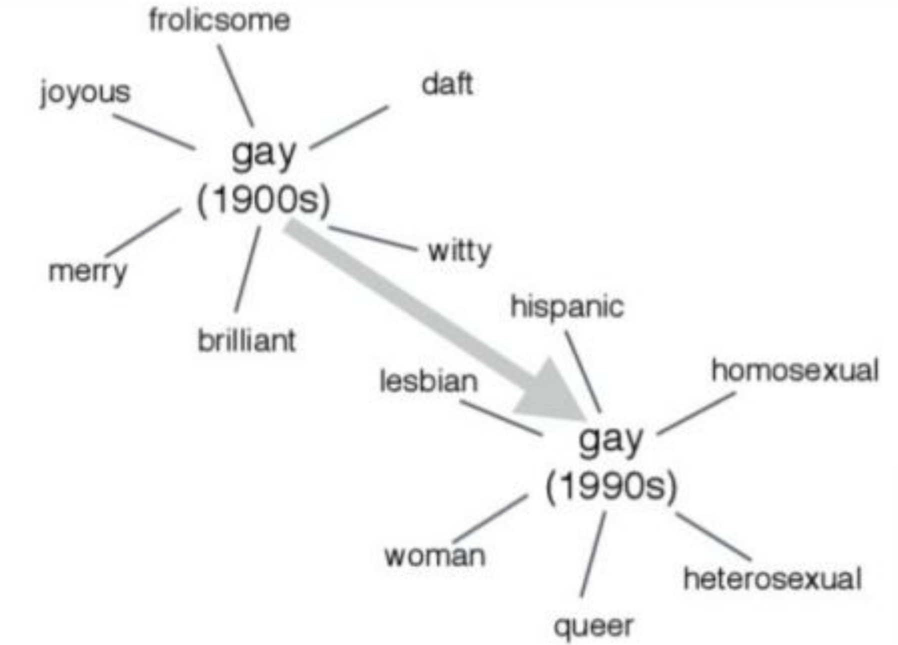
➔ She is a walking encyclopedia.

➔ He broke my heart.

➔ A great movie for a sunday nap!

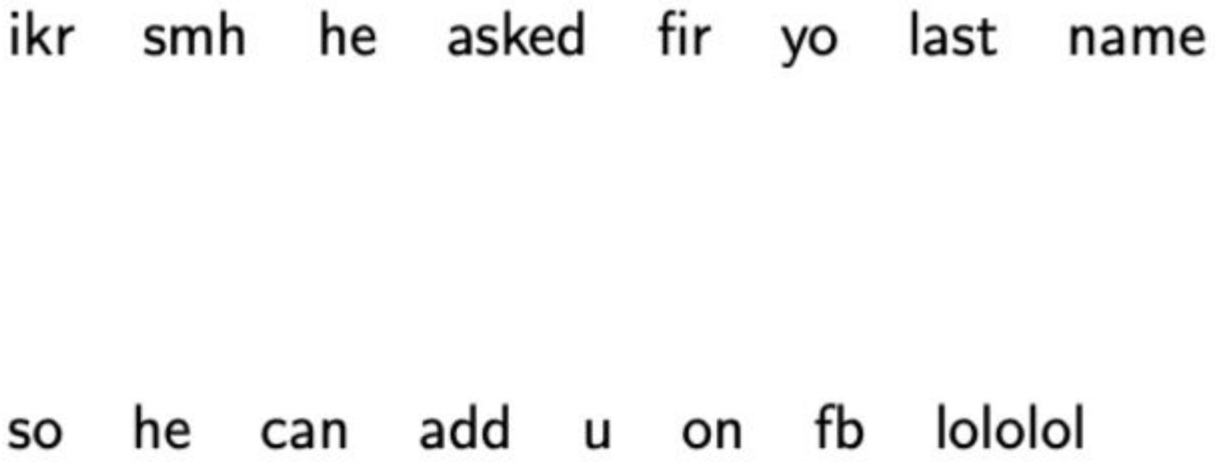


Language keeps evolving



Meaning changes

Language keeps evolving….

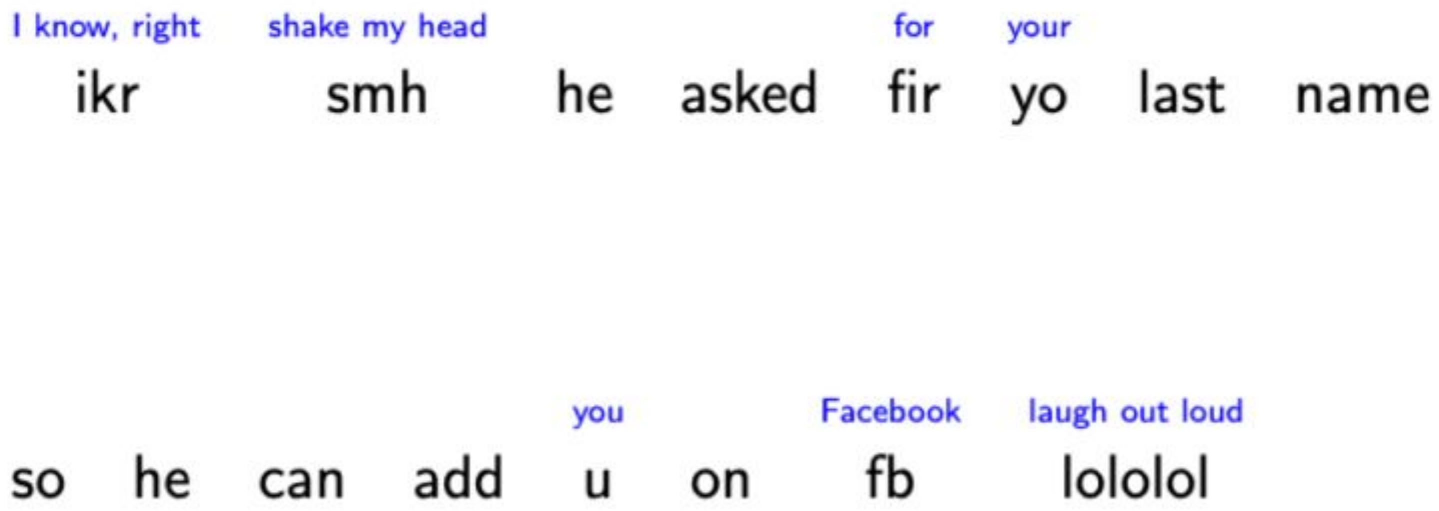


ikr smh he asked fir yo last name

so he can add u on fb lololol

New words are introduced

Language keeps evolving….

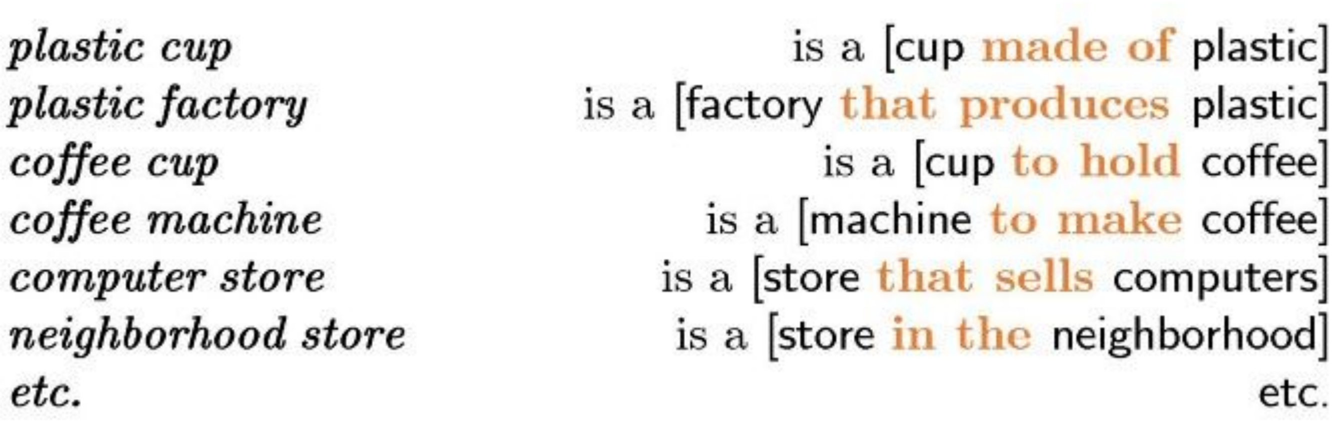


I know, right	shake my head			for	your		
ikr	smh	he	asked	fir	yo	last	name

				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol

Linguistic communication is compressed

- We all know what we all know, but machine don’t!! (Missing Text Phenomenon)

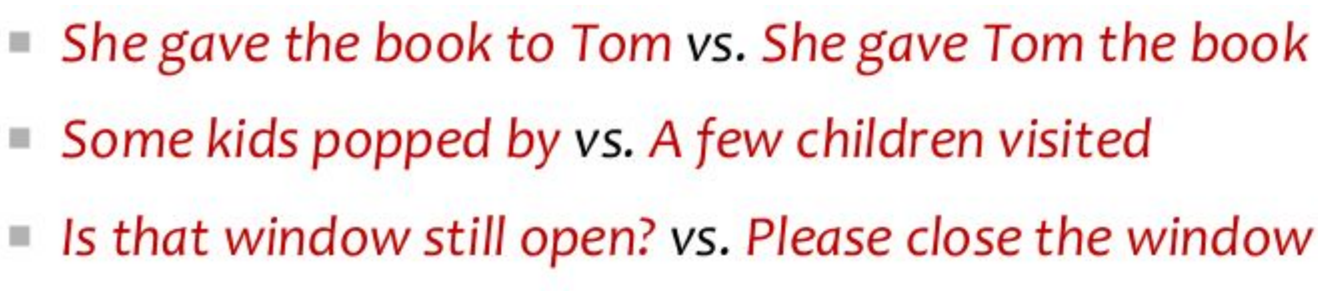


plastic cup	is a [cup made of plastic]
plastic factory	is a [factory that produces plastic]
coffee cup	is a [cup to hold coffee]
coffee machine	is a [machine to make coffee]
computer store	is a [store that sells computers]
neighborhood store	is a [store in the neighborhood]
etc.	etc.

The Missing Text Phenomenon, Again: the case of Compound Nominals
https://medium.com/ontologik/the-missing-text-phenomenon-again-the-case-of-compound-nominals-2776ad81fe38

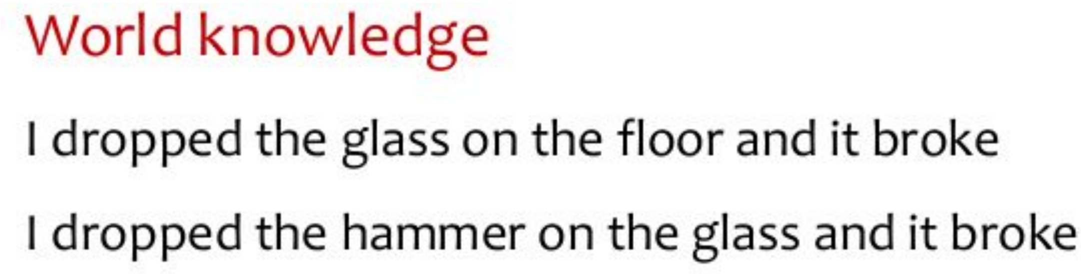
Expressivity

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:



- *She gave the book to Tom* vs. *She gave Tom the book*
- *Some kids popped by* vs. *A few children visited*
- *Is that window still open?* vs. *Please close the window*

Unmodeled Variables

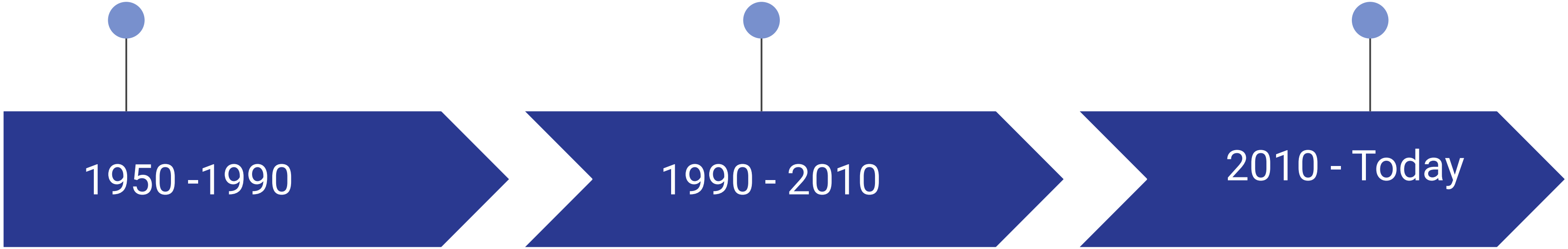


World knowledge

I dropped the glass on the floor and it broke

I dropped the hammer on the glass and it broke

Evolution of NLP over the years



Rule based NLP

1950 -1990

- Uses a well-defined set of rules
- Not scalable
- Need domain expertise

Statistical NLP

1990 - 2010

- Uses statistics to learn rules
- Relies on hand-crafted features
- Need domain expertise

Neural NLP

2010 - Today

- Uses artificial neural networks
- Extracts features from raw data
- No need of domain expertise

Rule - Based NLP System

Think of some useful rules to filter spam emails ??

Rule - Based NLP System

Think of some useful rules to filter spam emails ??

R1: misspelling in company names
R2: overuse of words like “lottery”,”dollars”
R3: Unnecessary urgency
R4: Inappropriate languages
R5:...

Statistical NLP System

- Relies on probability and statistics to learn rules or pattern from data

Eg:

“...won 5 million dollars...bank account…….click” ✉

What is the *__probability of finding these words__* in a spam vs genuine email?

Check whether **P(spam|email) > P (genuine|email)** ?

Statistical NLP System

- Auto completion can be done the same way!!!

I love chocolate ____________

What of these words is most probable to occur next in the above sentence?

(Pizza , dog , van, Cake , drink , pencil)

P(Cake|chocolate) > P(drink |chocolate) > P(pizza |chocolate) …..

Statistical NLP System

- Auto completion can be done the same way!!!

I love chocolate ____________

What of these words is most probable to occur next in the above sentence?

(Pizza , dog , van, Cake , drink , pencil)

P(Cake|chocolate) > P(drink |chocolate) > P(pizza |chocolate) …..

Statistical NLP System

You shall know a word by the company it keeps (Firth, J. R. 1957:11).

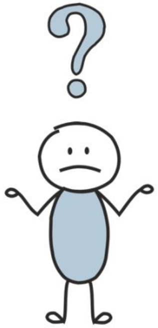
A bottle of tezgüino is on the table.

Everyone likes tezgüino.

Tezgüino makes you drunk.

We make tezgüino out of corn.

Can you understand what tezgüino means ?



Statistical NLP System

You shall know a word by the company it keeps (Firth, J. R. 1957:11).

(1) A bottle of _______ is on the table.

(2) Everyone likes _______ .

(3) _______ makes you drunk.

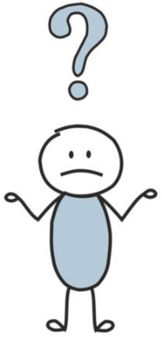
(4) We make _______ out of corn.

What other words fit into these contexts ?

	(1)	(2)	(3)	(4)	...
tezgüino	1	1	1	1	
loud	0	0	0	0	
motor oil	1	0	0	1	
tortillas	0	1	0	1	
wine	1	1	1	0	

← contexts

← rows show contextual properties: 1 if a word can appear in the context, 0 if not



Statistical NLP System

You shall know a word by the company it keeps (Firth, J. R. 1957:11).

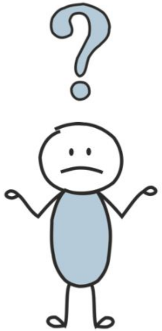
(1) A bottle of _______ is on the table.

(2) Everyone likes _______ .

(3) _______ makes you drunk.

(4) We make _______ out of corn.

	(1)	(2)	(3)	(4)	…
tezgüino	1	1	1	1	
loud	0	0	0	0	
motor oil	1	0	0	1	
tortillas	0	1	0	1	
wine	1	1	1	0	



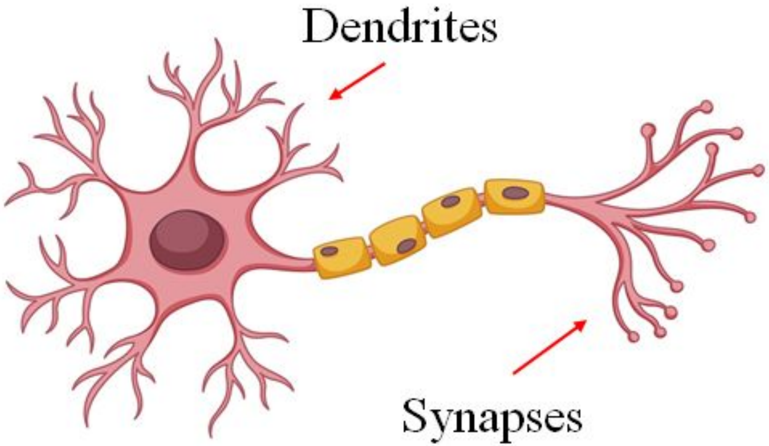
rows are similar → meanings of the words are similar

Is this true?

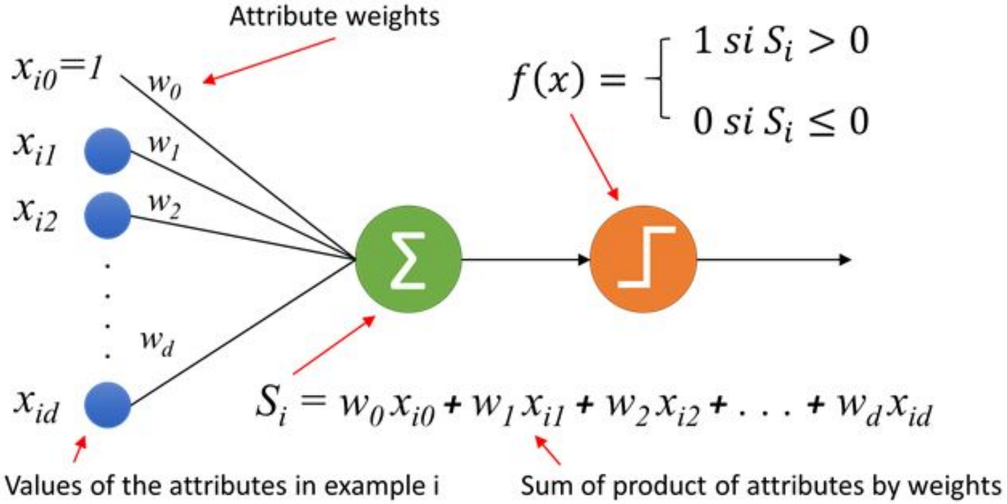
Neural NLP System

Artificial Neural Networks

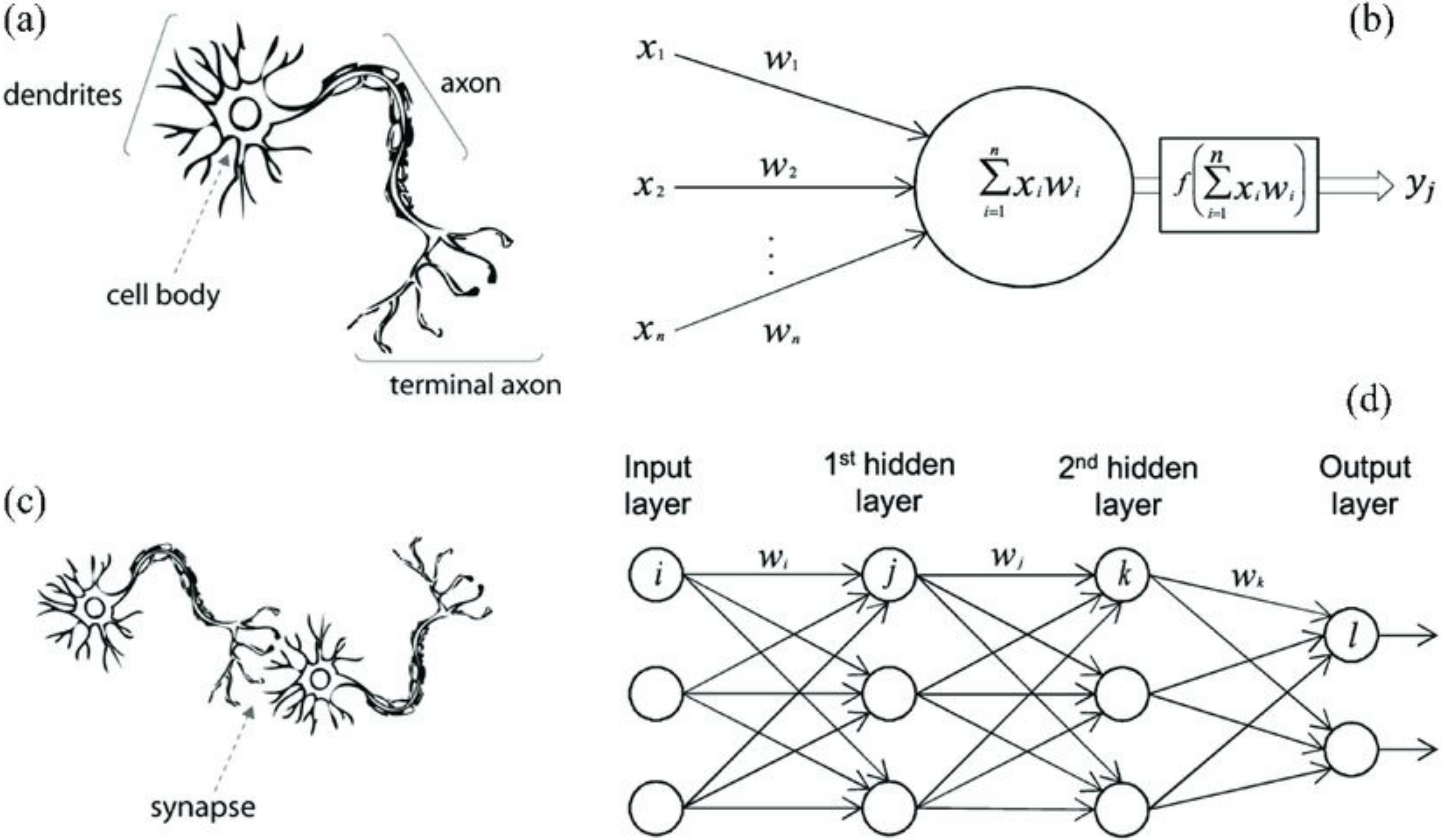
- Inspired by neurons in human brain



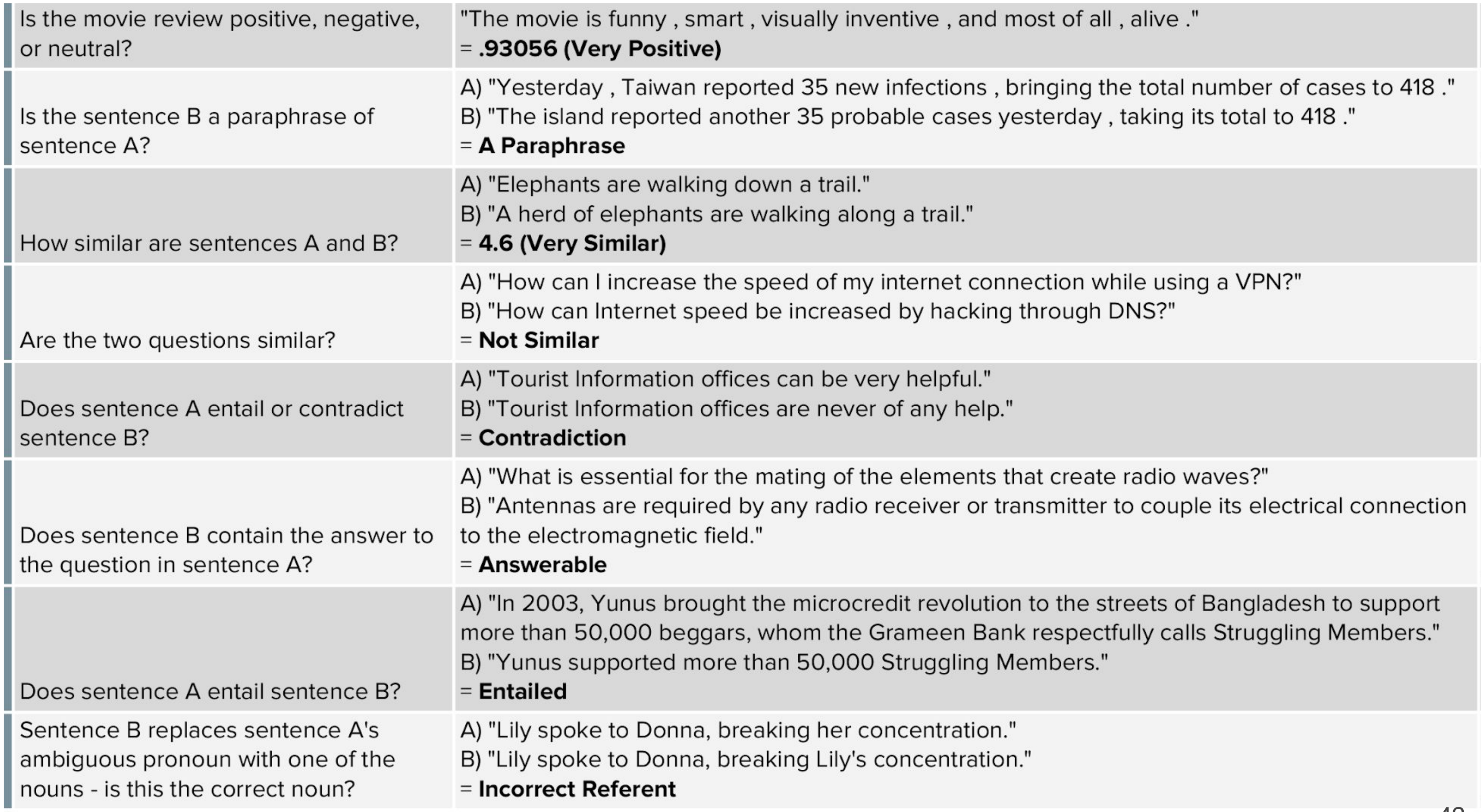
NEURON


Attribute weights
$x_{i0}=1$
x_{i1}
x_{i2}
x_{id}
w_0
w_1
w_2
w_d
Σ
$f(x) = \begin{cases} 1 \; si \; S_i > 0 \\ 0 \; si \; S_i \leq 0 \end{cases}$
$S_i = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id}$
Values of the attributes in example i
Sum of product of attributes by weights

PERCEPTRON


(a)
dendrites
axon
cell body
terminal axon
(b)
x_1
w_1
x_2
w_2
x_n
w_n
$\sum_{i=1}^{n} x_i w_i$
$f\left(\sum_{i=1}^{n} x_i w_i\right)$
y_j
(c)
synapse
(d)
Input layer
1st hidden layer
2nd hidden layer
Output layer
i
w_i
j
w_j
k
w_k
l

Classic NLU problems (still very relevant)



Sentiment Analysis	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**
Paraphrase Identification	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**
Similarity scoring	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**
Duplicate question	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**
Language Inference	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**
Question Answering	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**
Recognizing Textual Entailment	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**
Coreference Resolution	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**

Research/Project Ideas in Core NLP

Reading Comprehension
Visual Question Answering
Dialogue System
Event Extraction
Emotion Recognition
Semantic Parsing
Relational Reasoning
Abuse Detection
Stance Detection

Hate Speech Detection
Fake News Detection
Language Identification
Code Generation
Bias Detection
Intent Detection
Authorship Verification
Clickbait Detection

About the Course:

Course Objectives :

CO1 : Understand the fundamentals of written language processing

CO2: Applying theses fundamentals in real world problems like POS tagging, Corpus development, WordNet, Dialogue processing, document retrieval, Machine translation etc etc

CO3: Creating resources for less resource languages

CO4: Case study of various typical Language processing tools.

Thank You!!