Artificial Intelligence Foundations and Applications

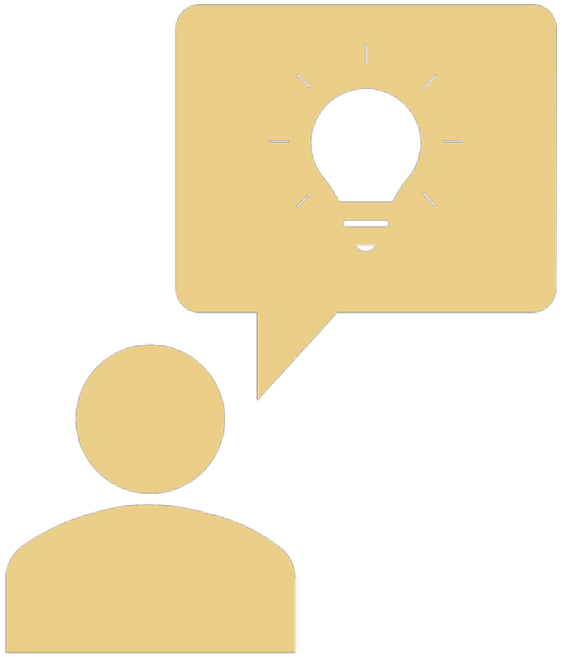
Introduction to Natural Language Processing

Centre of Excellence in Artificial Intelligence
IIT Kharagpur

Language is the Tool for Communication

Language is the Vehicle for

- Learning knowledge
- Transmitting information
- Expressing thoughts, perceptions, feelings, information
- Making sense of complex and abstract thought

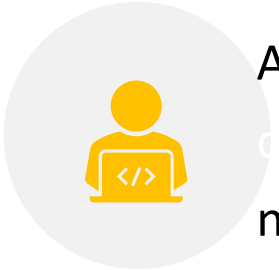


Communication is two-way

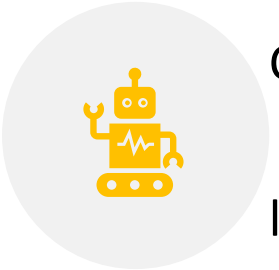
- Convey own ideas
- Receive thought of others

Natural Language Processing

Building computational systems for analyzing and understanding human language input and/or producing natural language output.



Allow computers to communicate with people using natural language.



Computational methods for understanding of human language.

- Automating Language
  - Analysis Language → Representation
  - Generation Representation → Language
  - Acquisition Obtaining the representation and necessary algorithms, from knowledge and data

Important Skills

Interact with our world using natural language

- E.g., Conversational agents

- Have computers read all the text out there
  - Retrieve
  - Answer questions
  - Summarize
  - Find new insights, Intelligence

Some Applications

- Search
- Language Translation
- Chatbots
- Question Answering
- Text Summarization
- Sentiment Analysis
- Topic Extraction
- Named Entity Recognition
- Relation Extraction
- Social Media Monitoring

Some application domains

Healthcare

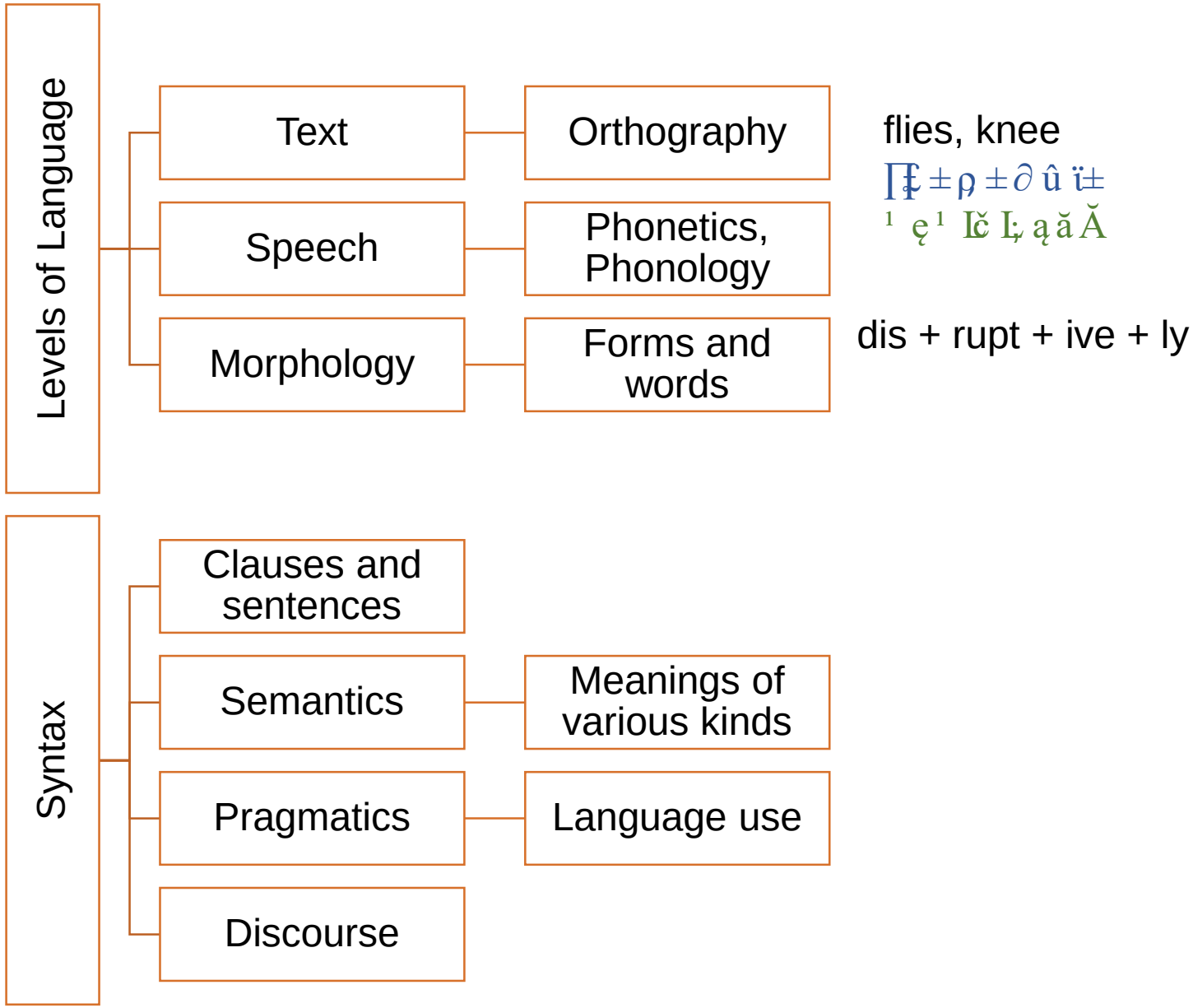
Education

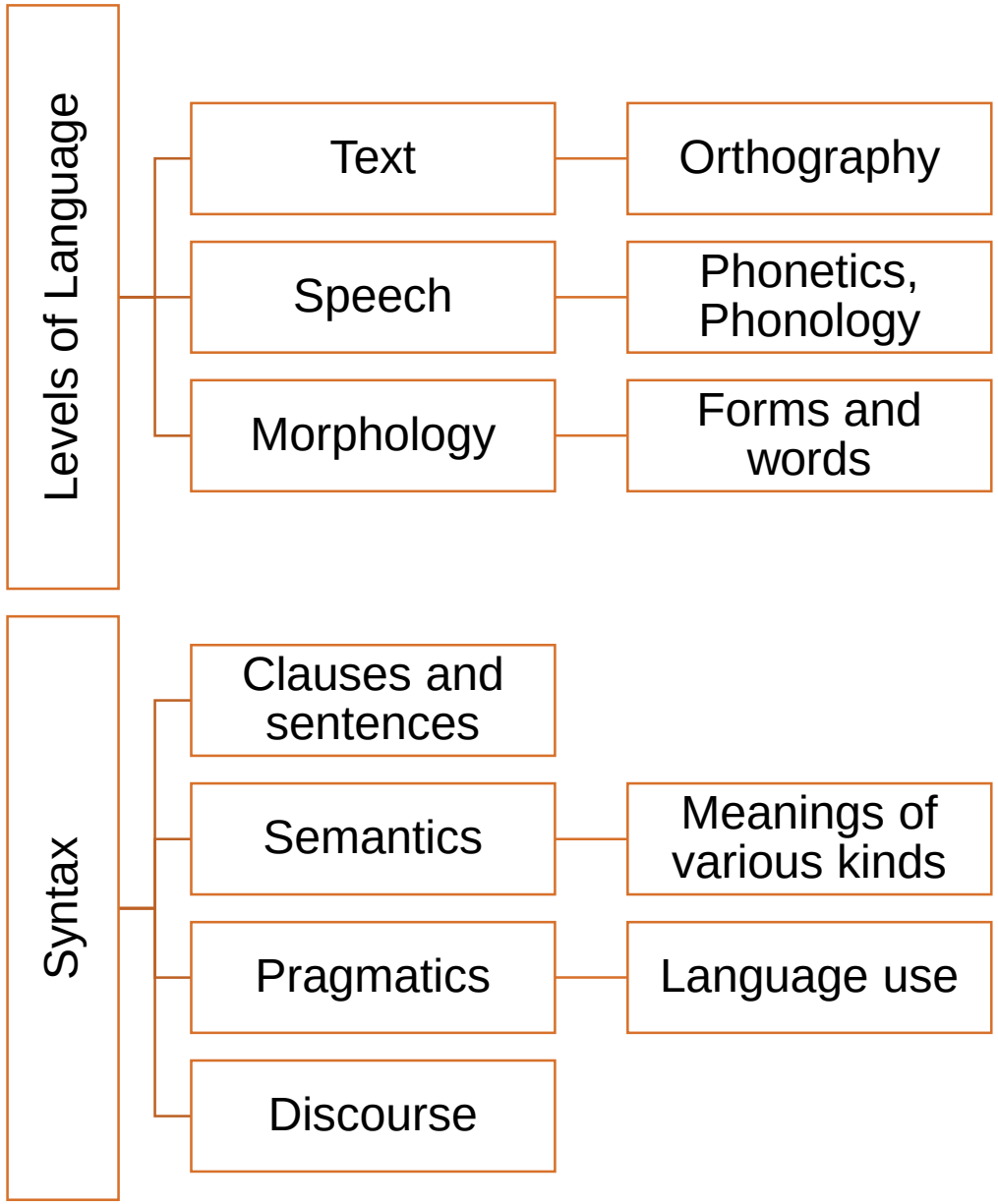
Agriculture

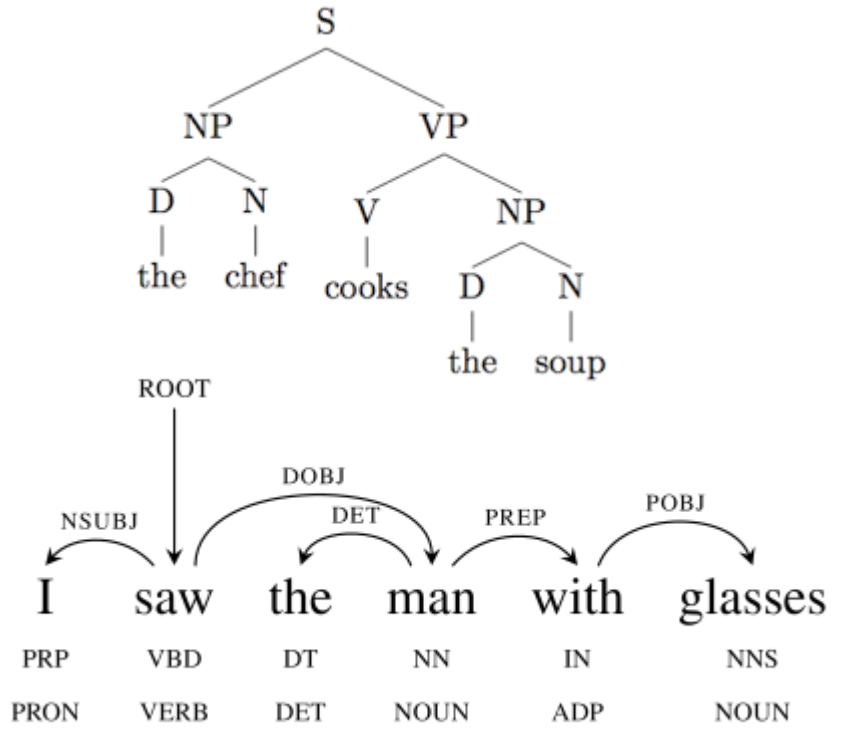
Law Enforcement

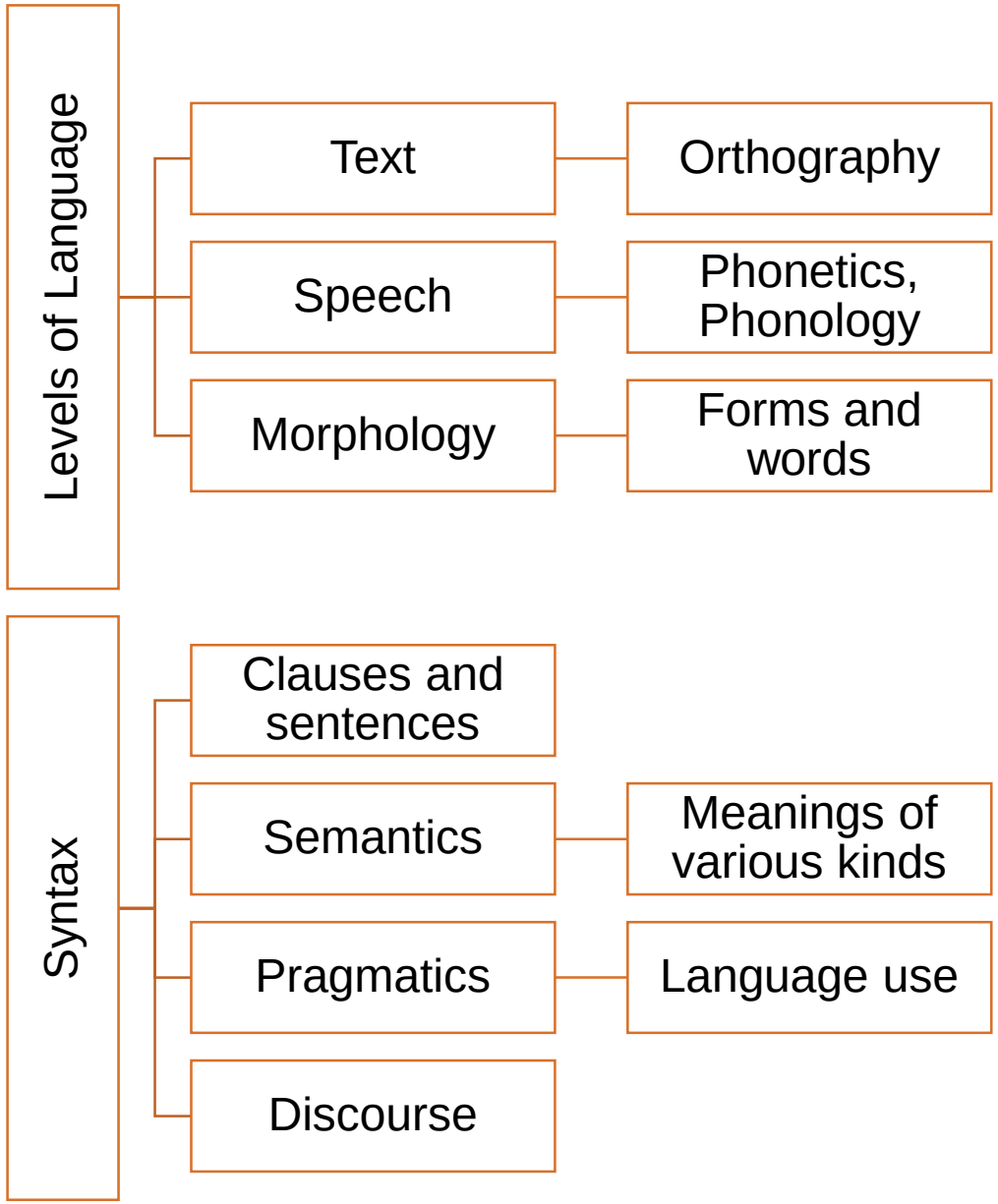
Analyze public feedback

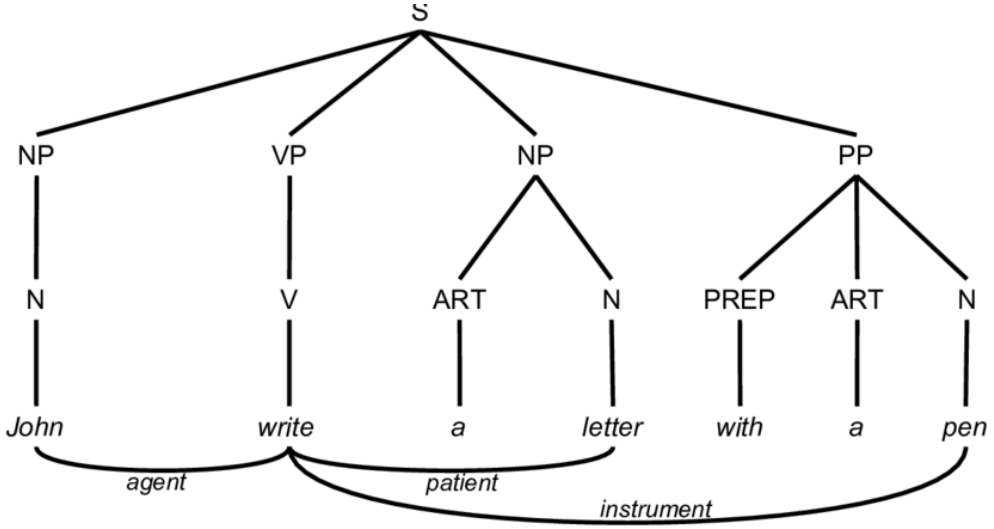
Improve Regulatory Compliance

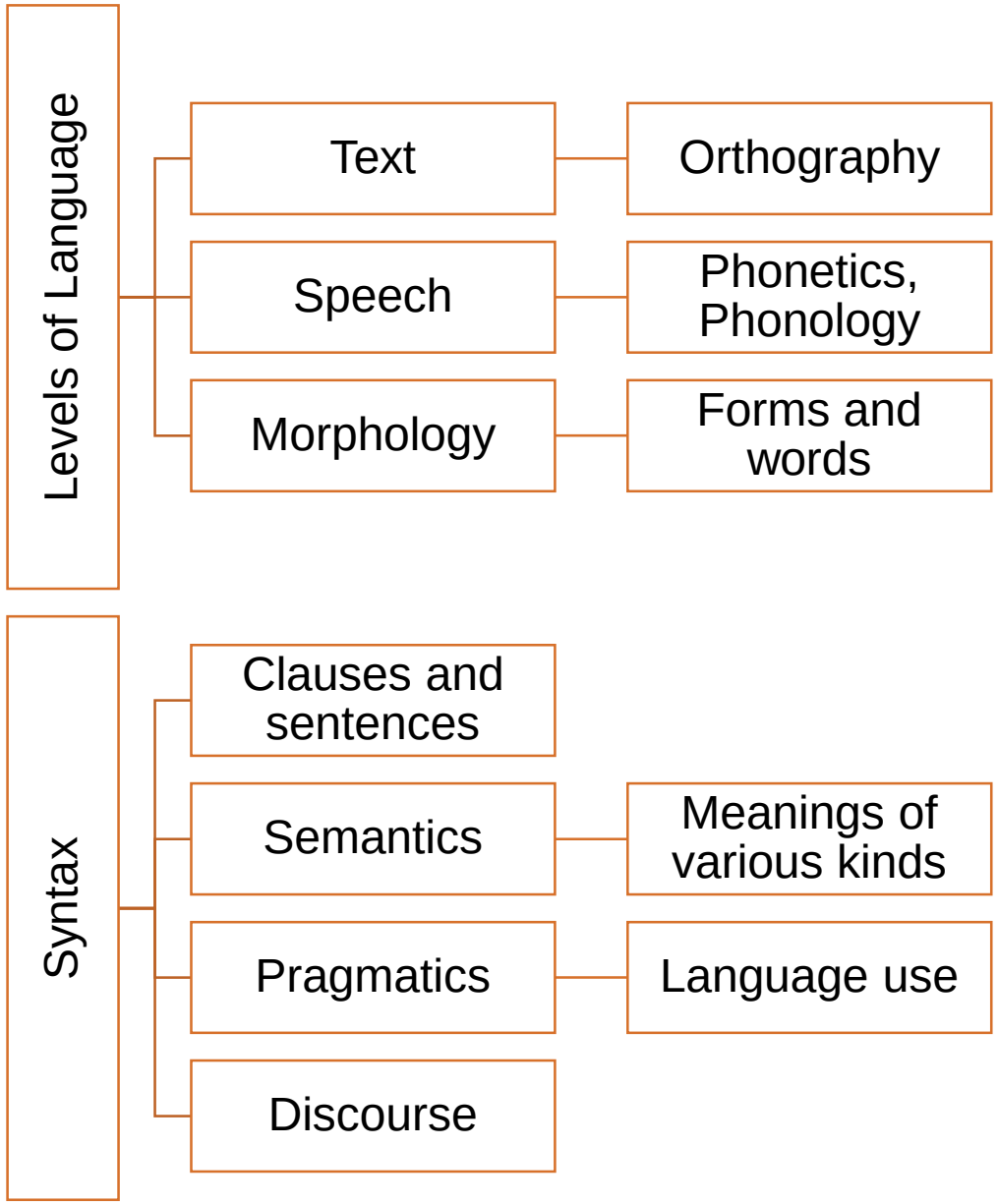

Levels of Language
Text
Orthography
flies, knee
Speech
Phonetics, Phonology
Morphology
Forms and words
dis + rupt + ive + ly
Syntax
Clauses and sentences
Semantics
Meanings of various kinds
Pragmatics
Language use
Discourse


Levels of Language
Text
Orthography
Speech
Phonetics, Phonology
Morphology
Forms and words
Syntax
Clauses and sentences
Semantics
Meanings of various kinds
Pragmatics
Language use
Discourse

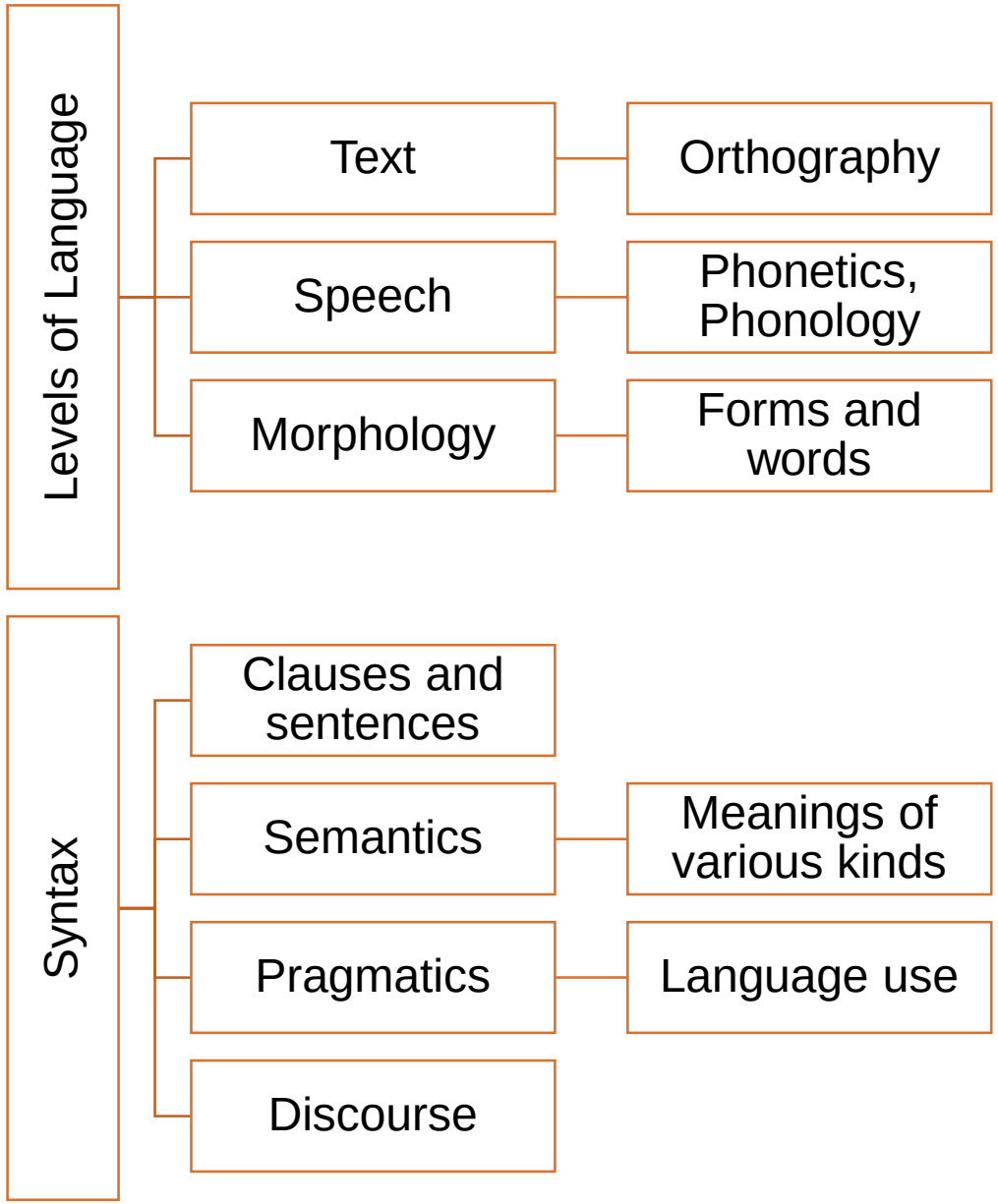

S
NP
VP
D
N
V
NP
the
chef
cooks
D
N
the
soup
ROOT
NSUBJ
DOBJ
DET
PREP
POBJ
I
saw
the
man
with
glasses
PRP
VBD
DT
NN
IN
NNS
PRON
VERB
DET
NOUN
ADP
NOUN


Levels of Language
Text
Orthography
Speech
Phonetics, Phonology
Morphology
Forms and words
Syntax
Clauses and sentences
Semantics
Meanings of various kinds
Pragmatics
Language use
Discourse


S
NP
VP
NP
PP
N
V
ART
N
PREP
ART
N
John
write
a
letter
with
a
pen
agent
patient
instrument


Levels of Language
Text
Orthography
Speech
Phonetics, Phonology
Morphology
Forms and words
Syntax
Clauses and sentences
Semantics
Meanings of various kinds
Pragmatics
Language use
Discourse

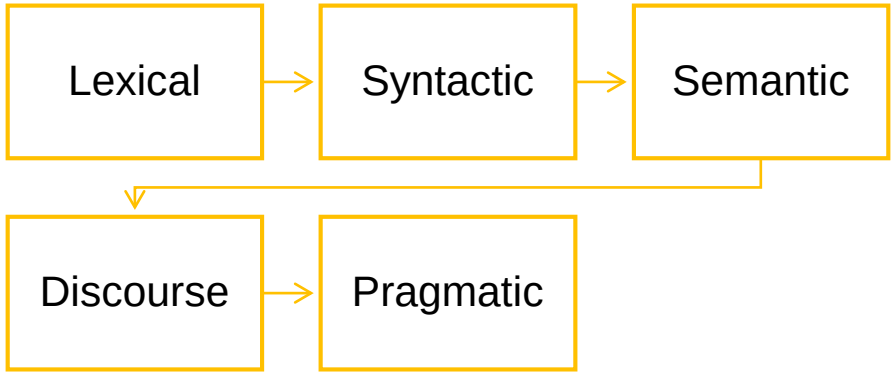

CECILIA,
CAN YOU STOP BY MY OFFICE? WE NEED TO TALK.
- PROF. JONES
"WE NEED TO TALK"
WHAT IT CAN MEAN:
A) YOU'RE FIRED.
B) I FORGOT WHAT YOUR PROJECT IS ABOUT. CAN YOU REMIND ME?
C) I HAVE A TRIVIAL REQUEST THAT I'M TOO EMBARRASSED TO WRITE IN AN EMAIL.
OH, SO THAT'S HOW YOU CHANGE THE LOGO!
I AM BOTH RELIEVED AND ANNOYED, PROF. JONES!
JORGE CHAM © 2013
WWW.PHDCOMICS.COM



This lesson explains **syntactic analysis**.
It discusses the algorithms for **this task**.

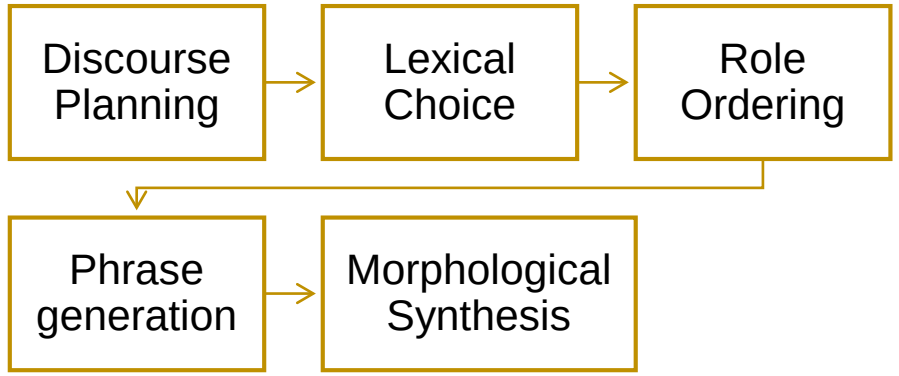
Analysis

Language → Representation



Generation

Representation → Language



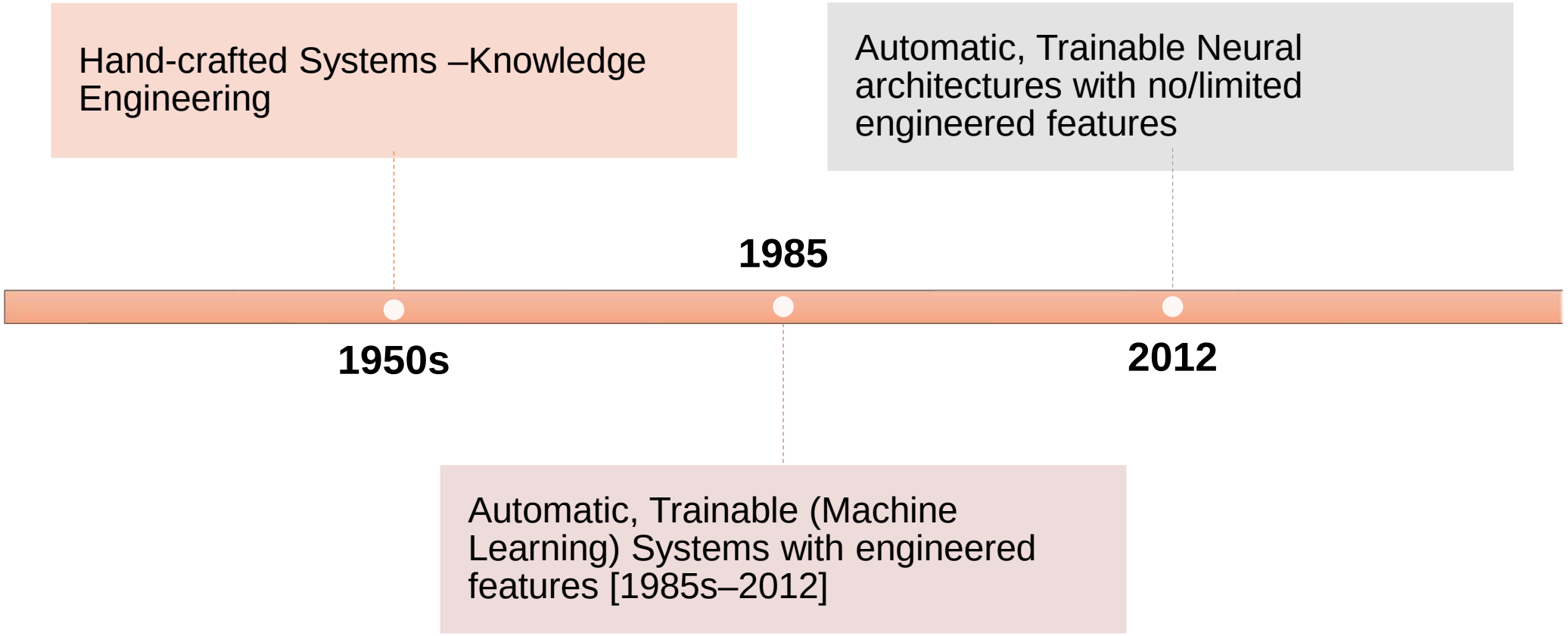
Acquisition

Obtaining the representation from knowledge and data

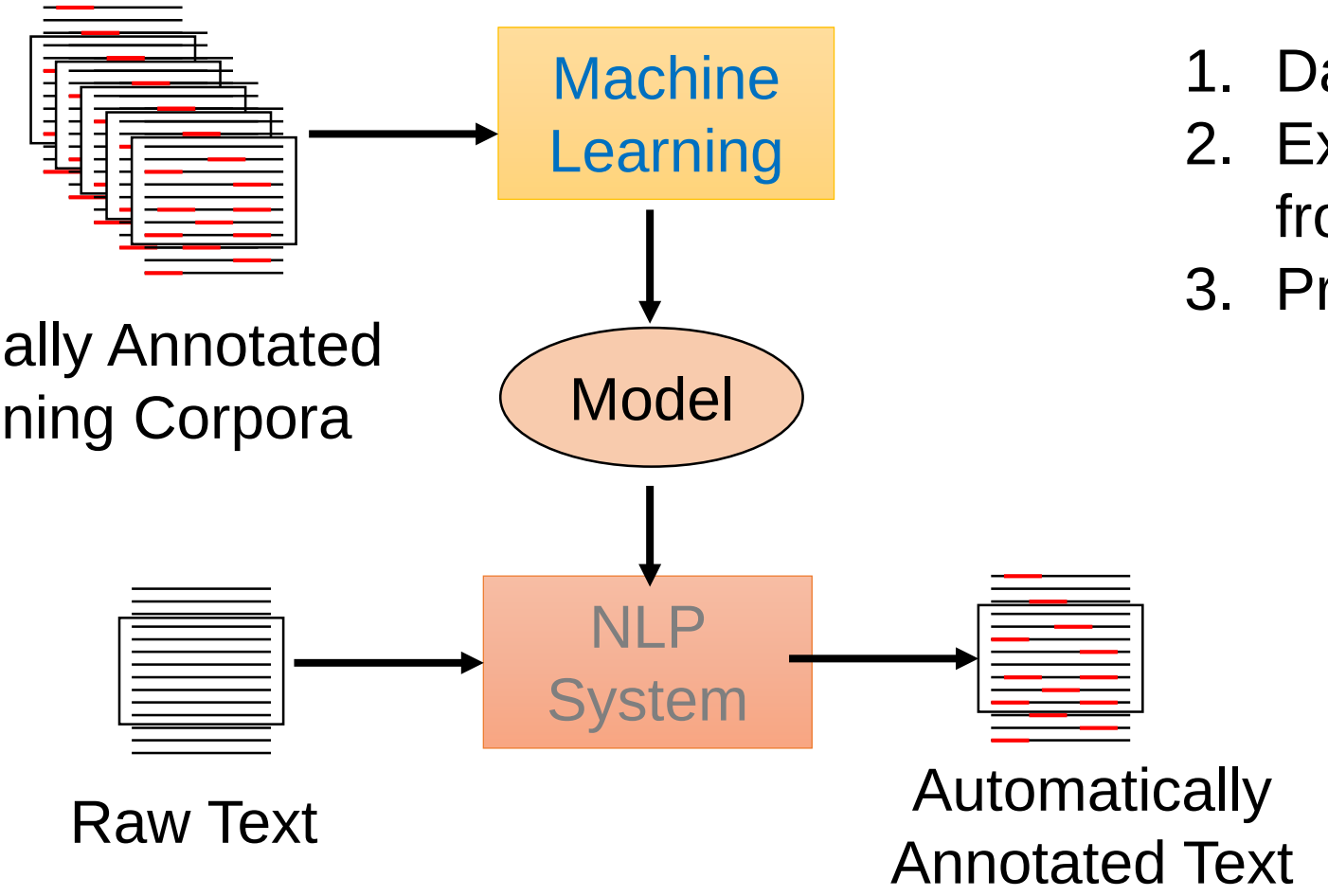
Hardness of NLP

- Ambiguity
- Richness (Variability)
  - Any meaning may be expressed many ways, and there are immeasurably many meanings.
- Linguistic diversity across languages, dialects, genres, styles

Three Generations of NLP

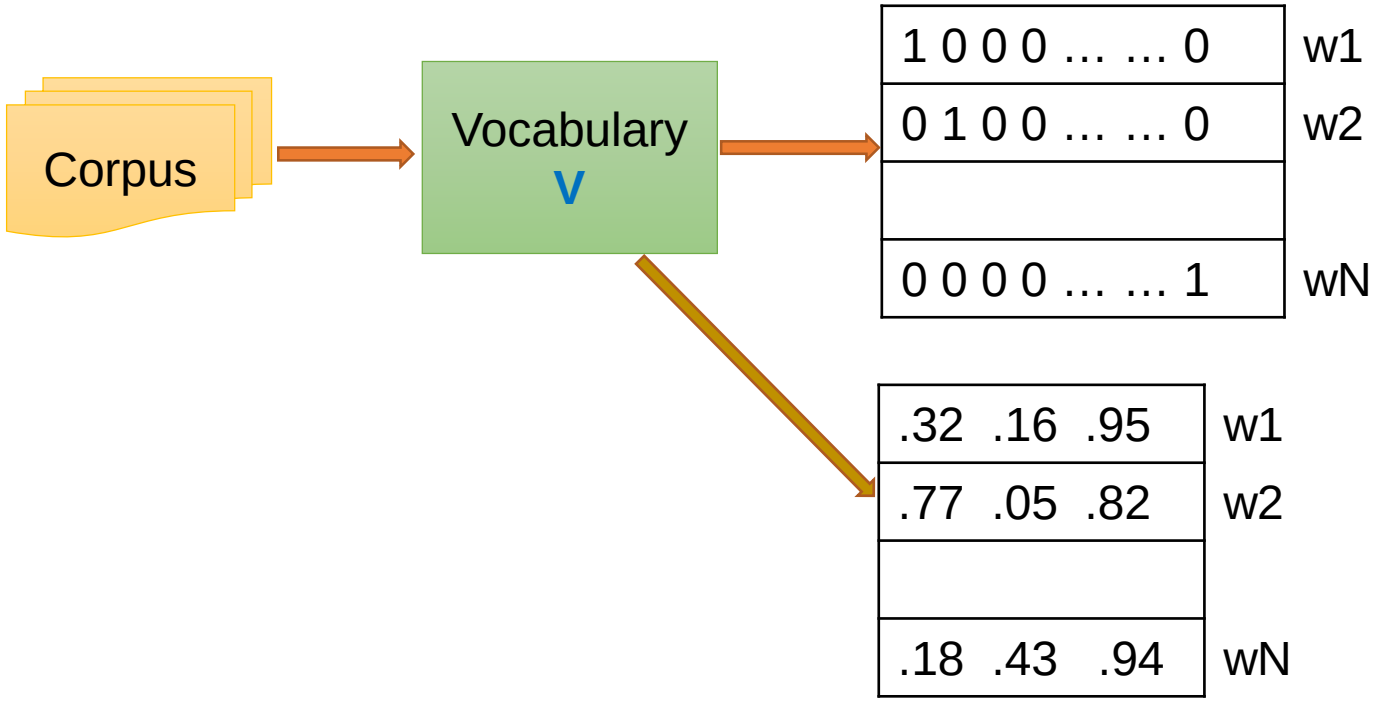


Machine Learning Approach to NLP



1. Data
2. Extraction of “features” from text data
3. Prediction of output

Word Representations



Ignores similarity between words

N=|V|

Word2vec: Represent each word with a low-dimensional dense vector

Model more generalizable

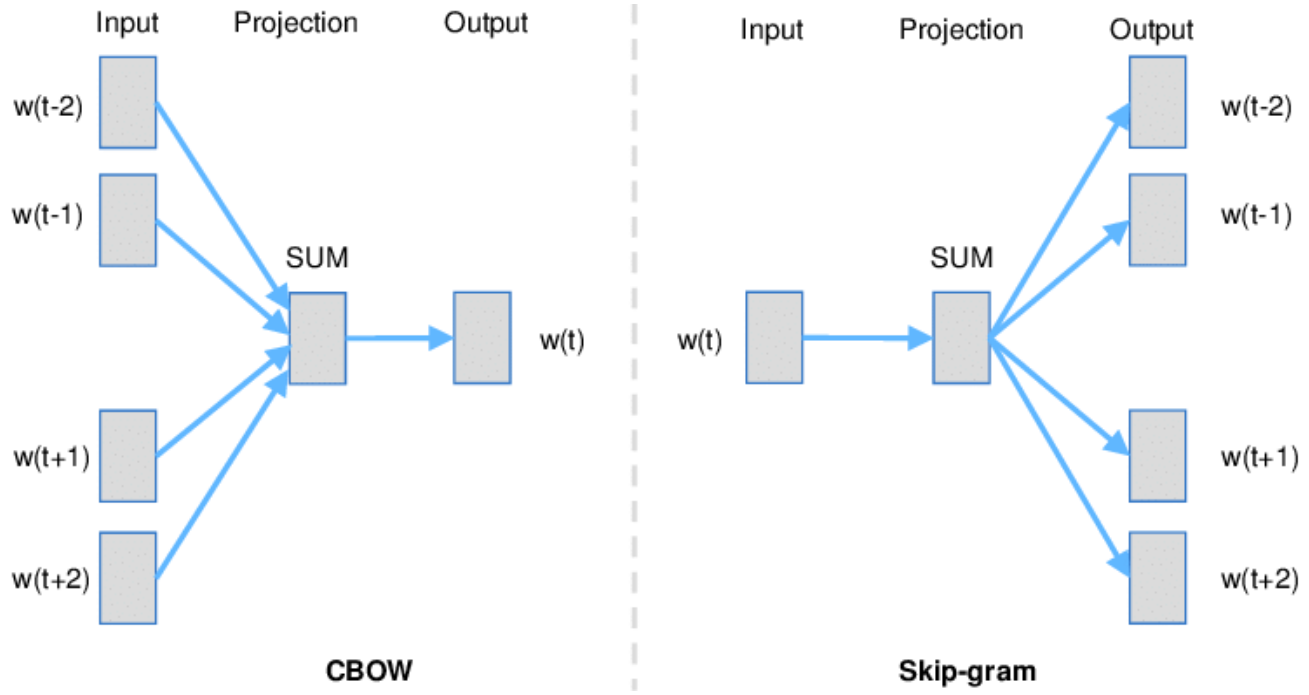
Word2vec Representations

"You shall know a word by the company it keeps"

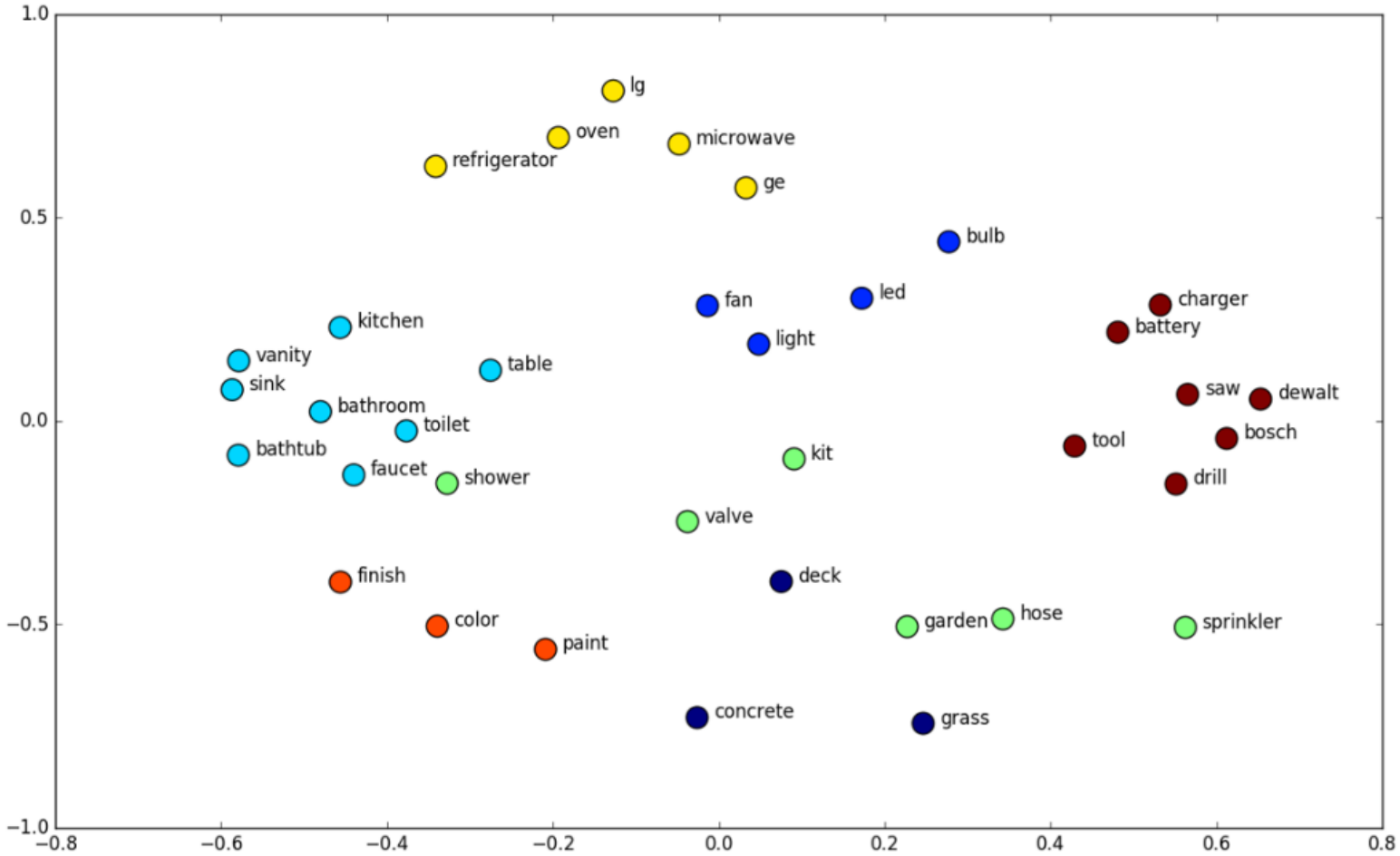
Key idea: Predict surrounding words of every word

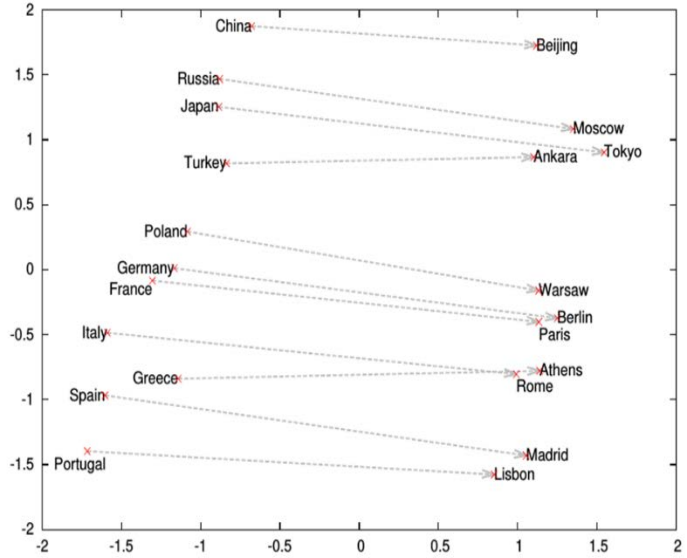
Assign each word a vector such that similar words have similar vectors

1. CBOW: P(Word|Context)
2. Skipgram: P(Context|Word)



Multilingual Embeddings





An interesting application

- Lawrence Berkeley lab material scientists applied word embedding to 3.3 million scientific abstracts published between 1922-2018.
  - 500k words. Vector size: 200 dimension,skip-gram model
- Captured things like periodic table and structure-property relationship in materials:

  ferromagnetic − NiFe + IrMn ≈ antiferromagnetic
- Discovered new thermoelectric materials

Nature, July 2019 V. Tshitonya et al, “Unsupervised word embeddings capture latent knowledge from materials science literature”.

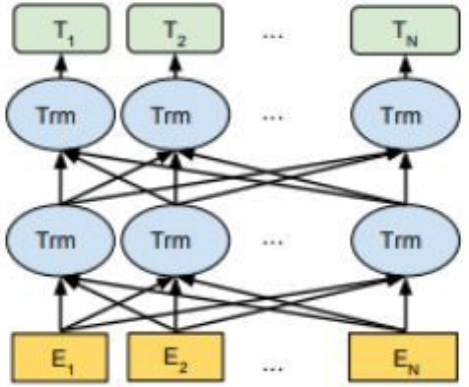
Contextualized Word Vectors

Incorporating context into word embeddings

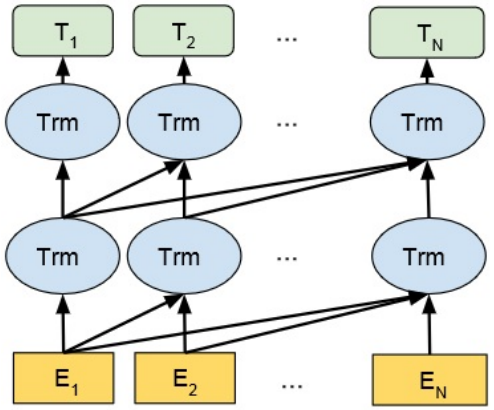
a watershed idea in NLP

- BERT: Bidirectional Encoder Representations from Transformers (BERT, 2018)
- GPT-2/3

Led to significant improvements on virtually every NLP task.



BERT Architecture



GPT

Language Models

How likely is a sentence $(w_1, w_2, \ldots, w_n)$?

- Predict the next word
- Complete the sentence

P (I saw a bus) >> P (eyes awe a boss)

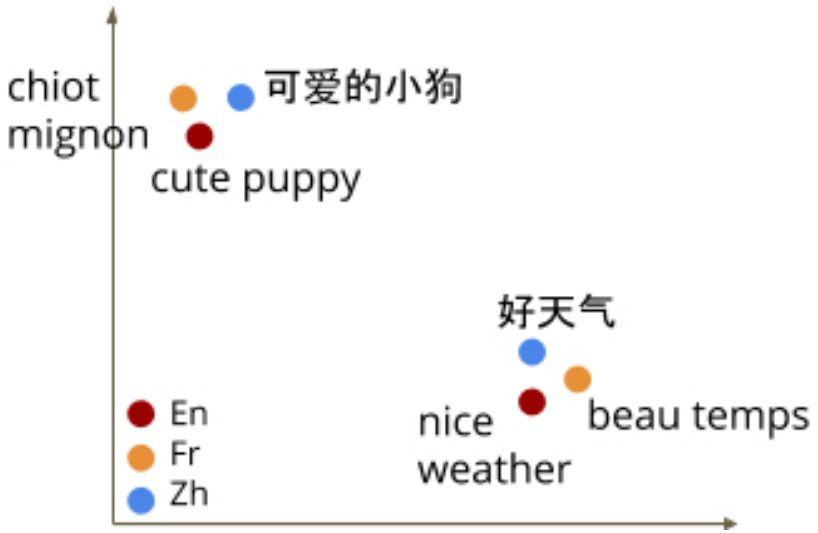
Pre-Trained Language Models

- Instead of training the model from scratch, you can use another pre-trained model as the basis and only fine-tune it to solve the specific NLP task.

Multilinguality

24 Aug 2020

- Google’s multilingual BERT model generates language-independent cross-language sentence embeddings for 109 languages



Machine Translation

Enabling access and communication in a multilingual world.

Breaking language barriers through machine translation (MT) is of the most important ways to bring people together

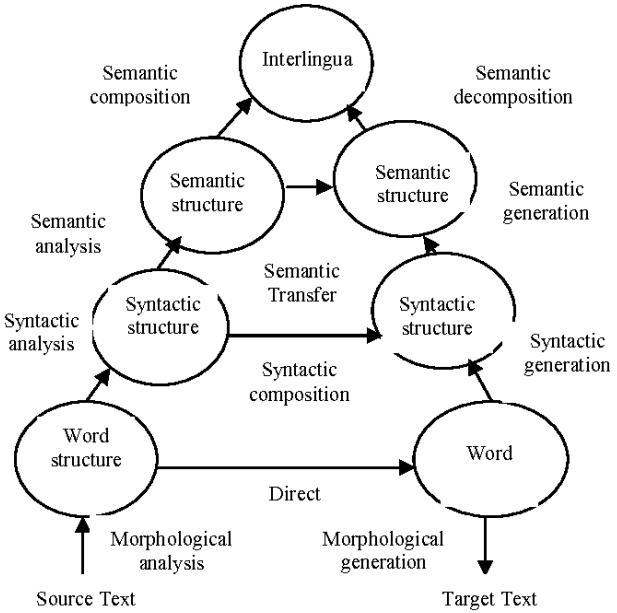
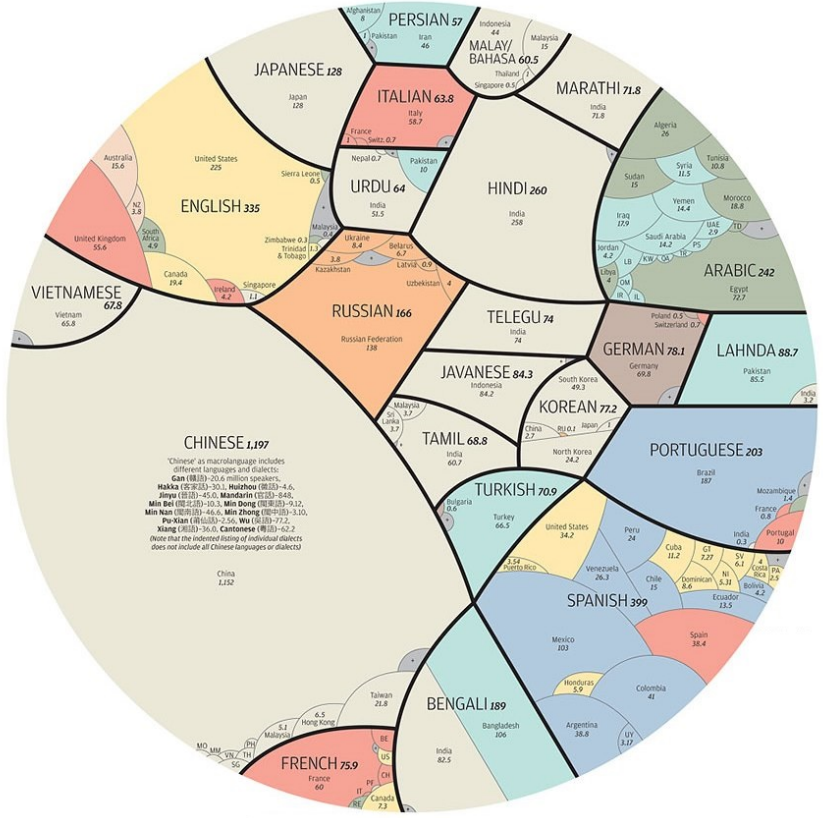
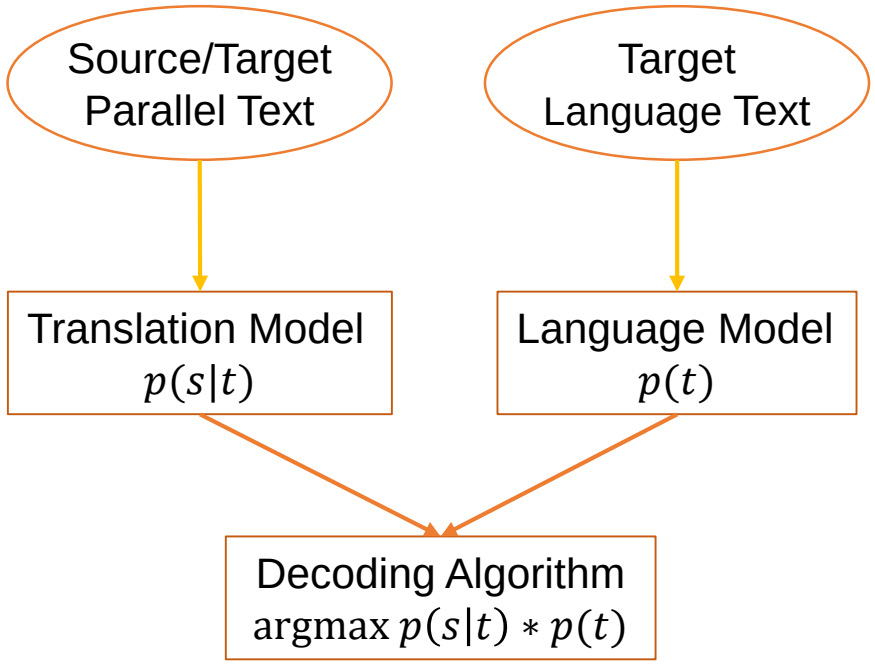


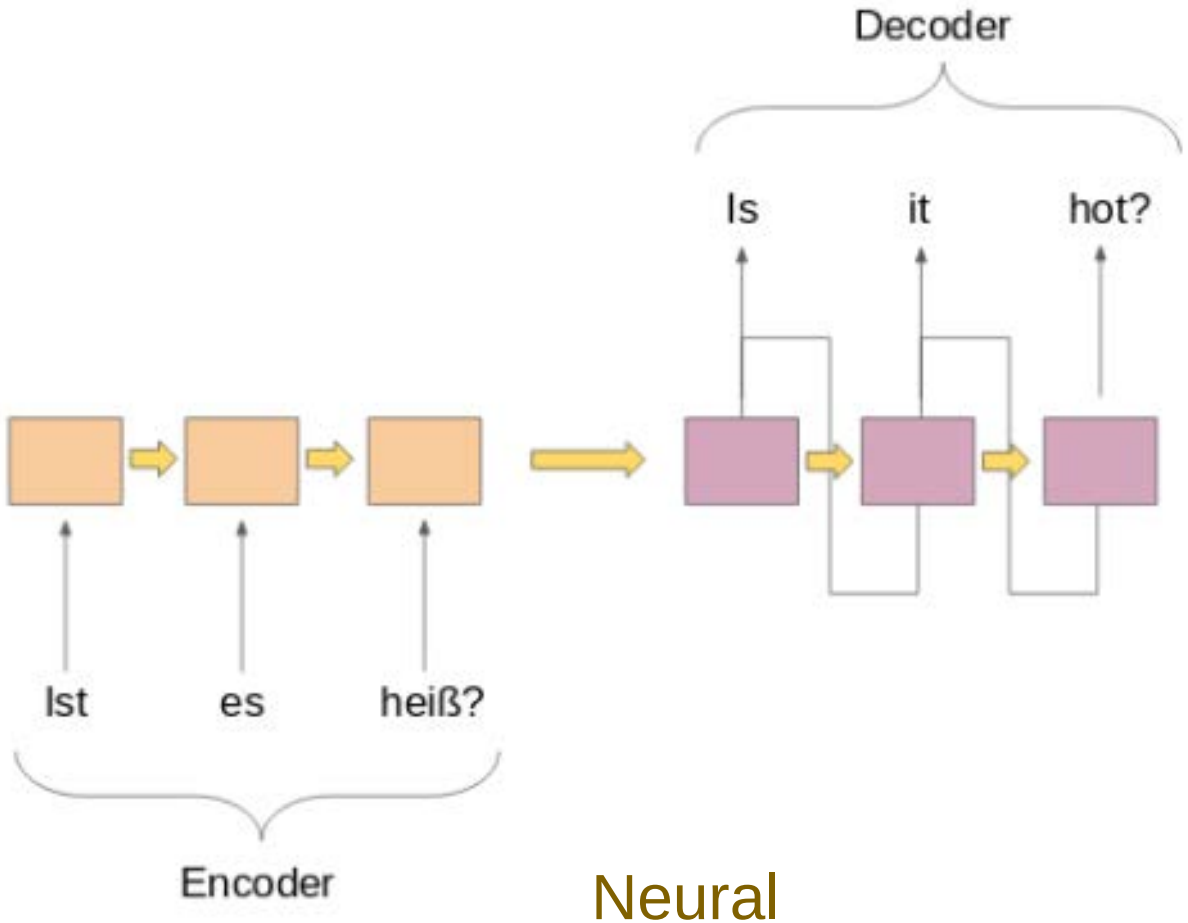
Figure 1. The Vauquois triangle.



Machine Translation



Statistical

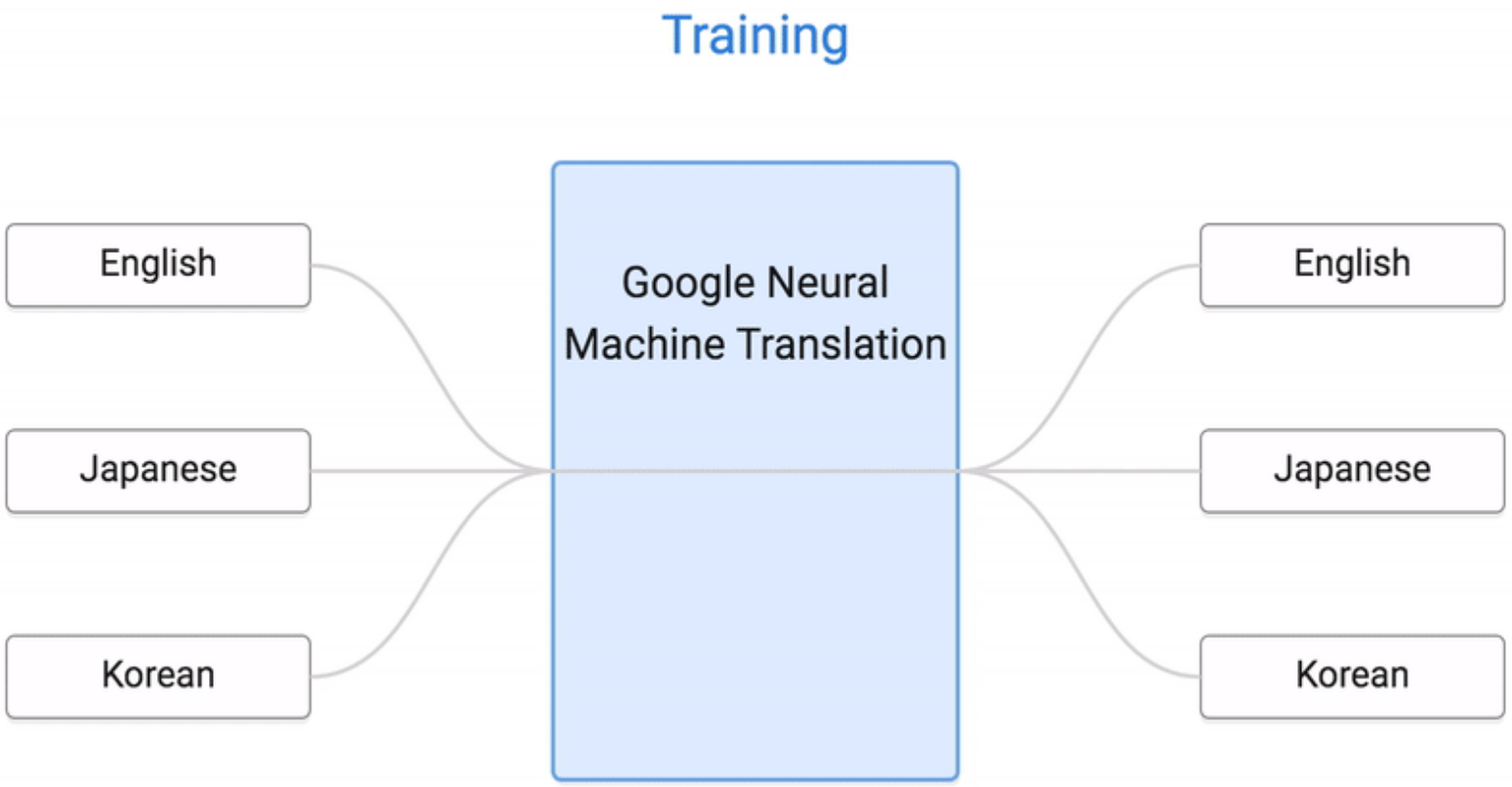


Neural

Challenge of Multilinguality

- Multilingual Representations
- Multilingual Models: Universal Models
  - One model to parse all languages
  - A universal model that can understand all
  - (Transfer learning, multilingual embeddings)

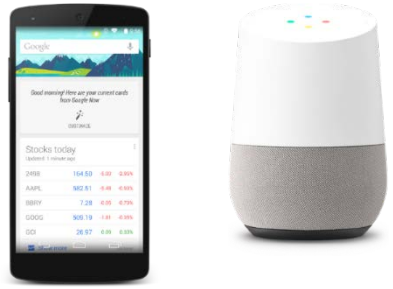
GNMT: Multilingual MT



Conversational Agents

- Personal Assistants
  - Alexa, SIRI, Cortana, Google Assistant
- Talking to your car
- Communicating with robots
- Clinical uses for mental health
- Chatbots
  - Customer Service, Call centres
  - Tutoring systems

Amazon Alexa/Echo (2014)

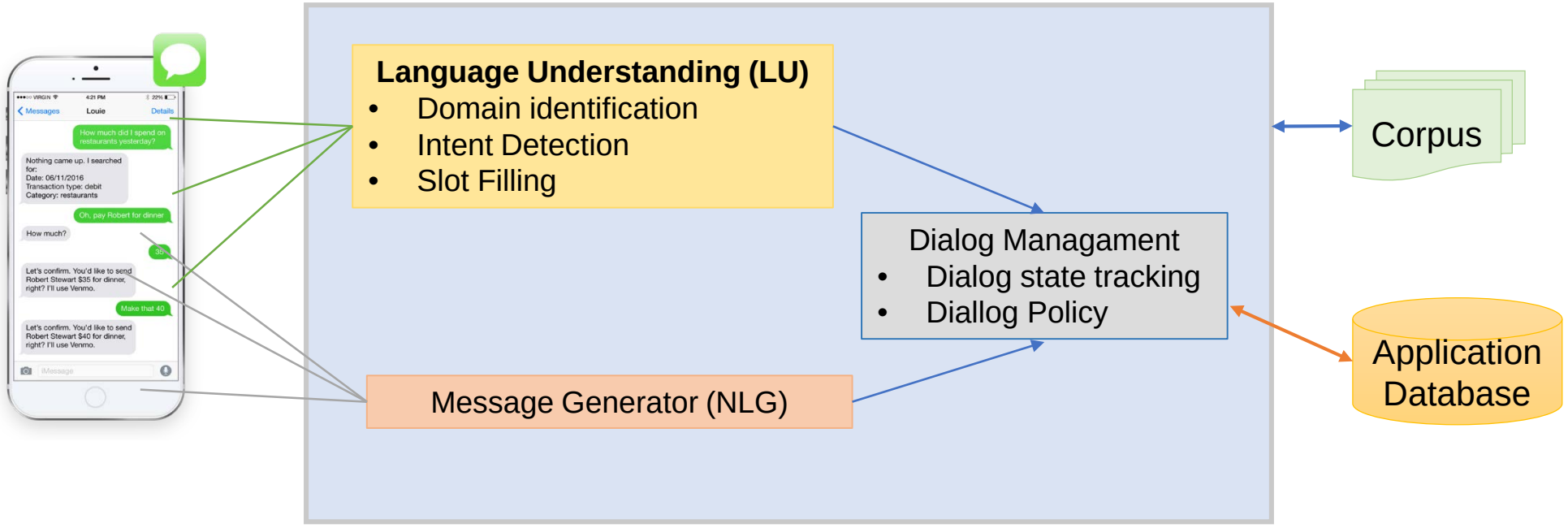


Microsoft Cortana (2014)

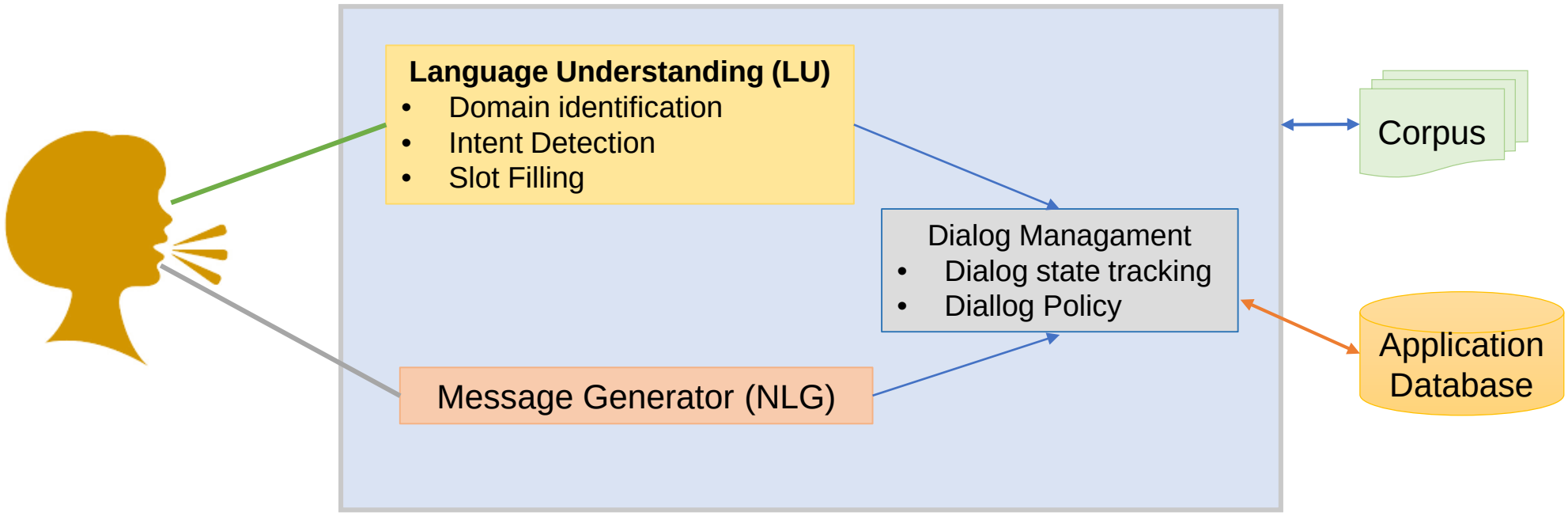
Chatbot:
Intention based agents

- Customer questions: analyzing the language
  - Identify user’s intent
- Formulate appropriate responses
  - NLG

Conversational Agents



Conversational Agents



Text Mining

Text Classification

- Text classification is the task of choosing correct class label for a given input.
  - Deciding whether an email is a spam or not (**spam detection**) .
  - Deciding whether the topic of a news article is from a fixed list of topic areas such as “sports”, “technology”, and “politics” (**document classification**).
  - Deciding whether a given occurrence of the word *bank* is used to refer to a river bank, a financial institution, the act of tilting to the side, or the act of depositing something in a financial institution (**word sense disambiguation**).

Named Entity Recognition

- Named entity refers to anything that can be referred to with a proper name.
- Named entity recognition aims to
  - Find spans of text that constitute proper names
  - Classify the entities being referred to according to their type

Type	Sample Categories	Example
People	Individuals, fictional Characters	***Turing*** is often considered to be the father of modern computer science.
Organization	Companies, parties	***Amazon*** plans to use drone copters for deliveries.
Location	Mountains, lakes, seas	The highest point in the ***Catalinas*** is ***Mount Lemmon*** at an elevation of 9,157 feet above sea level.
Geo-Political	Countries, states, provinces	The Catalinas, are located north, and northeast of ***Tucson***, ***Arizona***, ***United States.***
Facility	Bridges, airports	In the late 1940s, ***Chicago Midway*** was the busiest airport in the United States by total aircraft operations.
Vehicles	Planes, trains, cars	The updated ***Mini Cooper*** retains its charm and agility.

In practice, named entity recognition can be extended to types that are not in the table above, such as temporal expressions (time and dates), genes, proteins, medical related concepts (disease, treatment and medical events) and etc..

Named Entity Recognition

- Named entity recognition techniques can be categorized into knowledge-based approaches and machine learning based approaches.

Category	Advantage	Disadvantage	Tools /Ontology
Knowledge-based approach (rules & lexicons)	Require little training data	Creating lexicon manually is time-consuming and expensive; encoded knowledge might be importable across domains.	**General Entity Types**
			• WordNet
			• Lexicons created by experts
			Medical domain:
			• GATE (University of Sherfield)
			• UMLS (National library of Medicine)
			• MedLEE (Originally from Columbia University, commericalized now)
Machine learning approach - Conditional Random Field (CRF) - Hidden Markov Model (HMM)	Reduced human effort in maintaining rules and dictionaries	Prepared a set of annotated training data	**Conditional Random Field tools**
			• Stanford NER
			• CRF++
			• Mallet
			Hidden Markov Model tools
			• Mallet
			• Natural Language Toolkit(NLTK)

Entity Relation Extraction

- Entity relation extraction discerns the relationships that exist among the entities detected in a text. Entity relation extraction techniques are applied in a variety of areas.
  - Question Answering (e.g., IBM Watson)
    - Extracting entities and relational patterns for answering factoid question
  - Feature/Aspect based Sentiment Analysis
    - Extract relational patterns among entity, features and sentiments in text R(entity, feature, sentiment).
  - Mining bio-medical texts
    - Protein binding relations useful for drug discovery
    - Detection of gene-disease relations from biomedical literature
    - Finding drug-side effect relations in health social media

Sentiment Analysis

- Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source material.

- The rise of social media such as forums, micro blogging and blogs has fueled interest in sentiment analysis.
  - Online reviews, ratings and recommendations in social media sites have turned into a kind of virtual currency for businesses looking to market their products, identifying new opportunities and manage their reputations

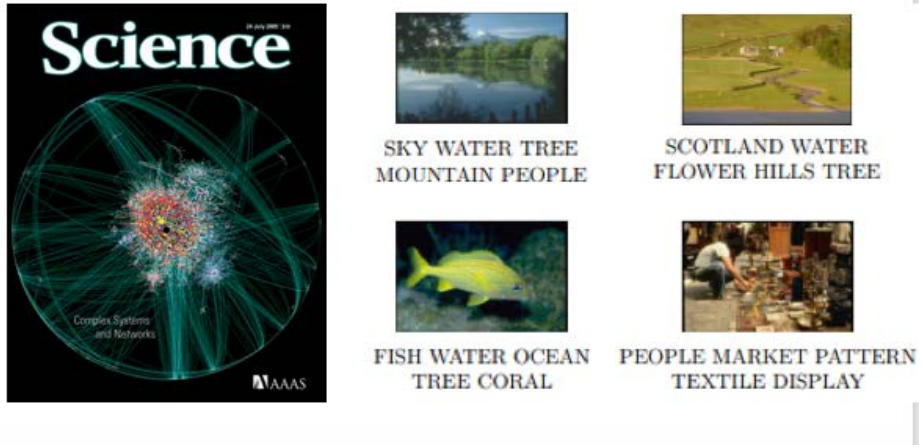
Sentiment Analysis

Task	description	Approaches	lexicons/ algorithms
Polarity Classification	**classifying a given text at the document, sentence, or feature/aspect level into positive, negative or neutral**	lexicon based scoring	SentiWordNet, LIWC
		machine learning classification	SVM
Affect Analysis	Classifying a given text into affect states such as "angry", "sad", and "happy"	lexicon based scoring	WordNet-Affect
		machine learning classification	SVM
Subjectivity Analysis	Classifying a given text into two classes: objective and subjective	lexicon based scoring	SentiWordNet, LIWC
		machine learning classification	SVM
Feature/Aspect Based Analysis	**Determining the opinions or sentiment expressed on different features or aspects of entities (e.g., the screen[feature] of a cell phone [entity])**	Named entity recognition + entity relation detection	SentiWordNet, LIWC, WordNet
			SVM
Opinion Holder /Target Analysis	Detecting the holder of a sentiment (i.e. the person who maintains that affective state) and the target (i.e. the entity about which the affect is felt)	Named entity recognition + entity relation detection	SentiWordNet, LIWC, WordNet
			SVM

Topic Modeling

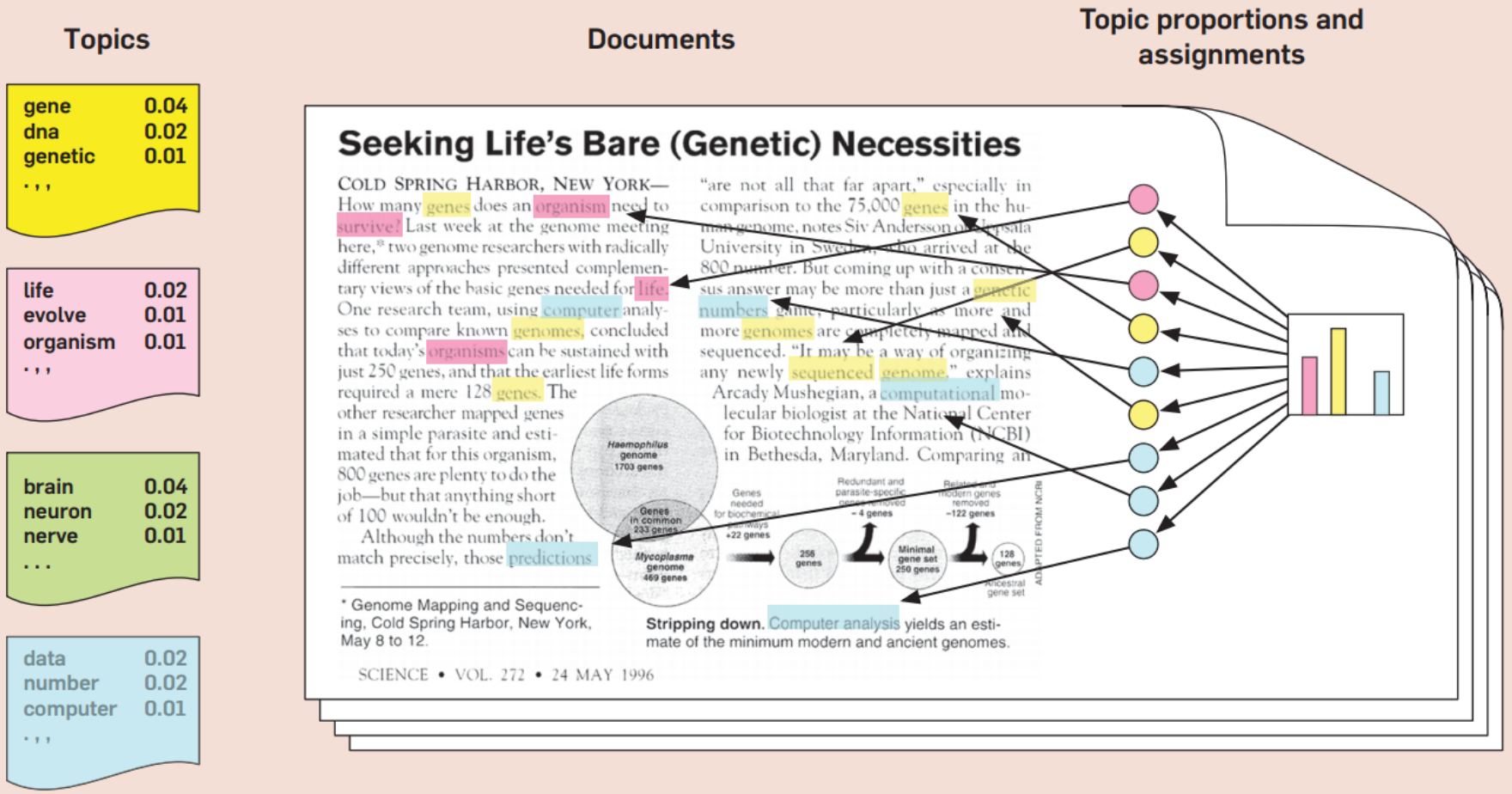
- Topic models : algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents.
  - Latent Dirichlet Allocation (LDA).

Topic modeling algorithms can be adapted to many kinds of data, e.g., **annotate documents and images; organize and browse large corpora; model topic evolution; categorize source code archives; discover influential articles**; etc.



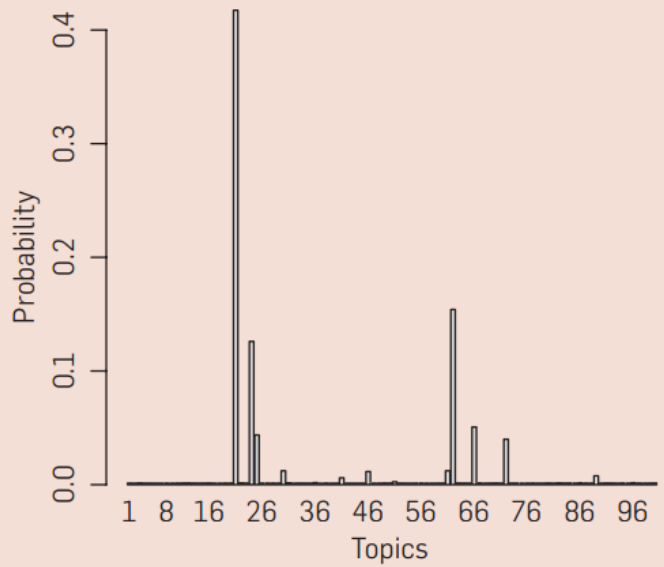
Topic Modeling - LDA

The figure below shows the intuitions behind **latent Dirichlet allocation.** We assume that some number of "topics", which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic .



Topic Modeling - LDA

The figure below show real inference with LDA. 100-topic LDA model is fitted to 17,000 articles from journal *Science*. At left are the inferred topic proportions for the example article in previous figure. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Information and Knowledge Extraction

Biomedical Text and Data Mining

- Drug Drug Interaction
  - Adverse Drug Reactions
- Drug Repurposing
  - the application of existing therapeutics to treat new disease indications.

J Clin Pharmacol. 2020 Feb 27. doi: 10.1002/jcph.1568. [Epub ahead of print]

Safety and Pharmacokinetics of DS-1040 Drug-Drug Interactions With Aspirin, Clopidogrel, and Enoxaparin.

Limsakun T[1], Dishy V[1], Mendell J[1], Pizzagalli F[2], Pav J[1], Kochan J[1], Vandell AG[1], Rambaran C[1], Kobayashi F[3], Orihashi Y[3], Warren V[1], McPhillips P[2], Zhou J[1].

Author information

Abstract

DS-1040, a novel low-molecular-weight inhibitor of activated thrombin-activatable fibrinolysis inhibitor, is under development for the treatment of thromboembolic diseases including venous thromboembolism and acute ischemic stroke. Here we describe the results of 3 studies that evaluated the safety and tolerability of DS-1040 along with the effect on DS-1040 pharmacokinetic (PK) parameters, when dosed alone or when coadministered with aspirin (NCT02071004), clopidogrel (NCT02560688), or enoxaparin in healthy subjects. Concomitant administration of single-dose DS-1040 with multiple-dose aspirin, multiple-dose clopidogrel, or single-dose enoxaparin, consistent with clinically relevant dose regimens, was safe and well tolerated with no serious treatment-emergent adverse events (TEAEs), TEAEs leading to discontinuation, bleeding-related TEAEs, and no significant changes in coagulation parameters. DS-1040 did not prolong bleeding time when administered concomitantly with aspirin or clopidogrel. In the aspirin study, DS-1040 PK was evaluated following the concomitant administration with multiple-dose aspirin, where the plasma DS-1040 exposure (peak plasma concentration [C_{max}] and area under the concentration-time curve [AUC_{inf}]) was to be similar to the data previously published in the first-in-human study of DS-1040 in healthy subjects. The PK parameters of DS-1040 coadministered with clopidogrel were similar to those of DS-1040 alone, with small increases in geometric means for C_{max} (7%) and AUC_{last} (9%). When coadministered with enoxaparin, the PK parameters of DS-1040 were not affected (1.1% and 1.5% decreases in geometric means for C_{max} and AUC_{last} , respectively). Therefore, concomitant administration of DS-1040 and clopidogrel or enoxaparin did not demonstrate PK drug-drug interactions.

KEYWORDS: drug-drug interactions; fibrinolysis inhibitor; pharmcokinetics; stroke; thrombosis

PMID: 32106339 DOI: 10.1002/jcph.1568

An abstract of a research article

Text (and Data) Mining based Inference

Resource Collection

Drug, Gene, Disease

Named Entity Recognition

Protein, Drugs

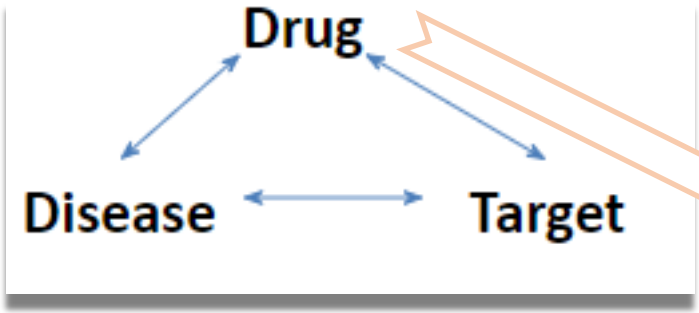
Coreference and Linking

Normalization

Disease

Relation Extraction

Intra sentence
PPI, DDI, Chem-Prot



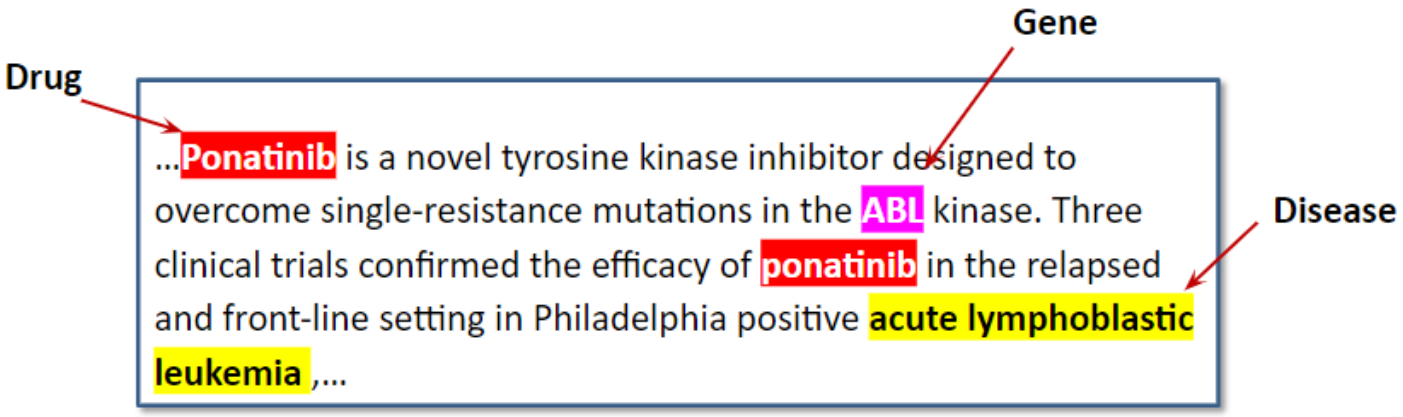
Task: Drug-Drug Interaction Prediction

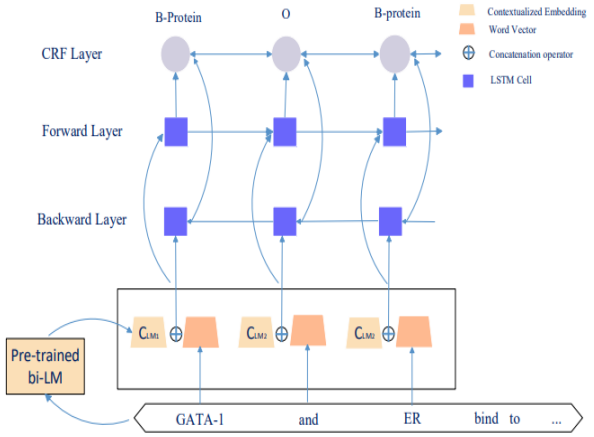
Knowledge Graph

Inference Algorithms

Biomedical Named Entity Recognition (NER)

Identification of *proper names* in texts, and **classification** into a set of predefined entity classess

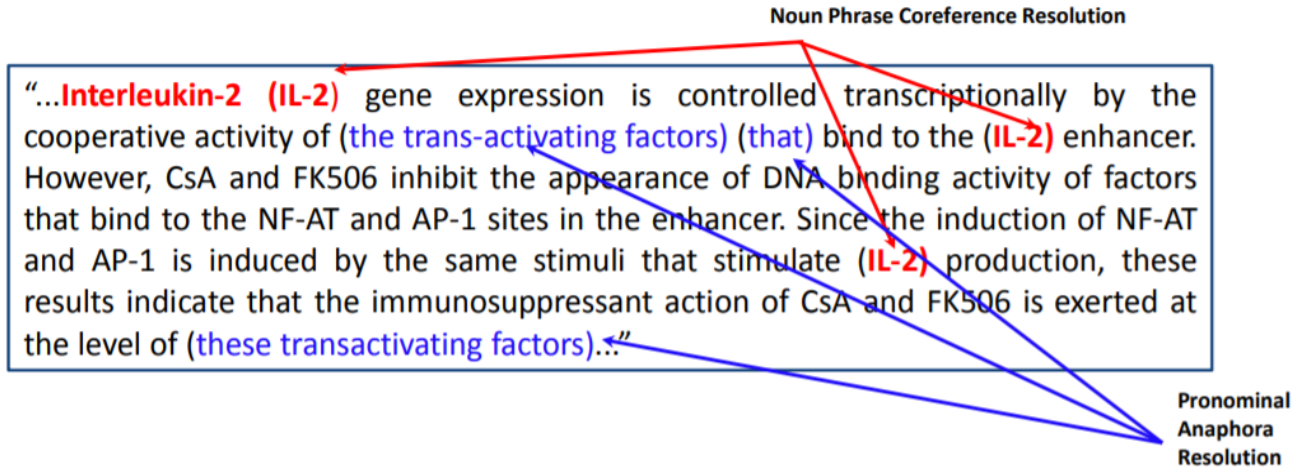




Biomedical Coreference Resolution

- The task of finding all the nominal and pronomial expressions that refer to the same entity in the text.

Example:

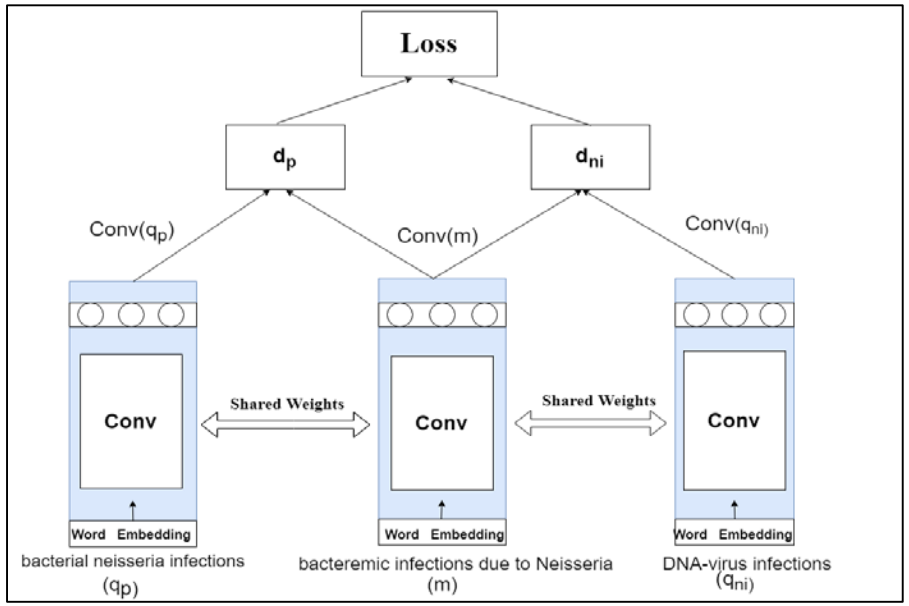


Biomedical Normalization

The task of mapping entities in the medical text to standard entities in a given Knowledge Base(KB).

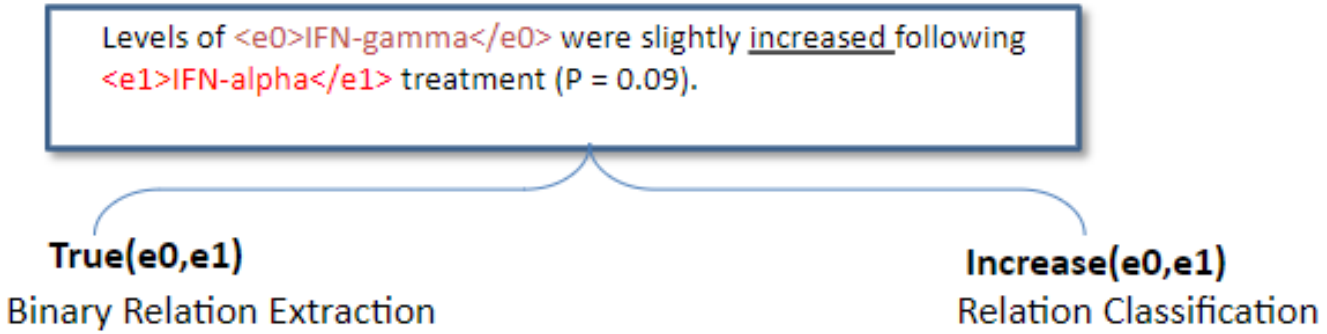
*"**Renal amyloidosis,** prevented by colchicine, is the most severe complication of FMF ..."*
Source : (PMID:10364520)

Knowledge Base ID (C538249) having synonyms like **Amyloidosis 8**.

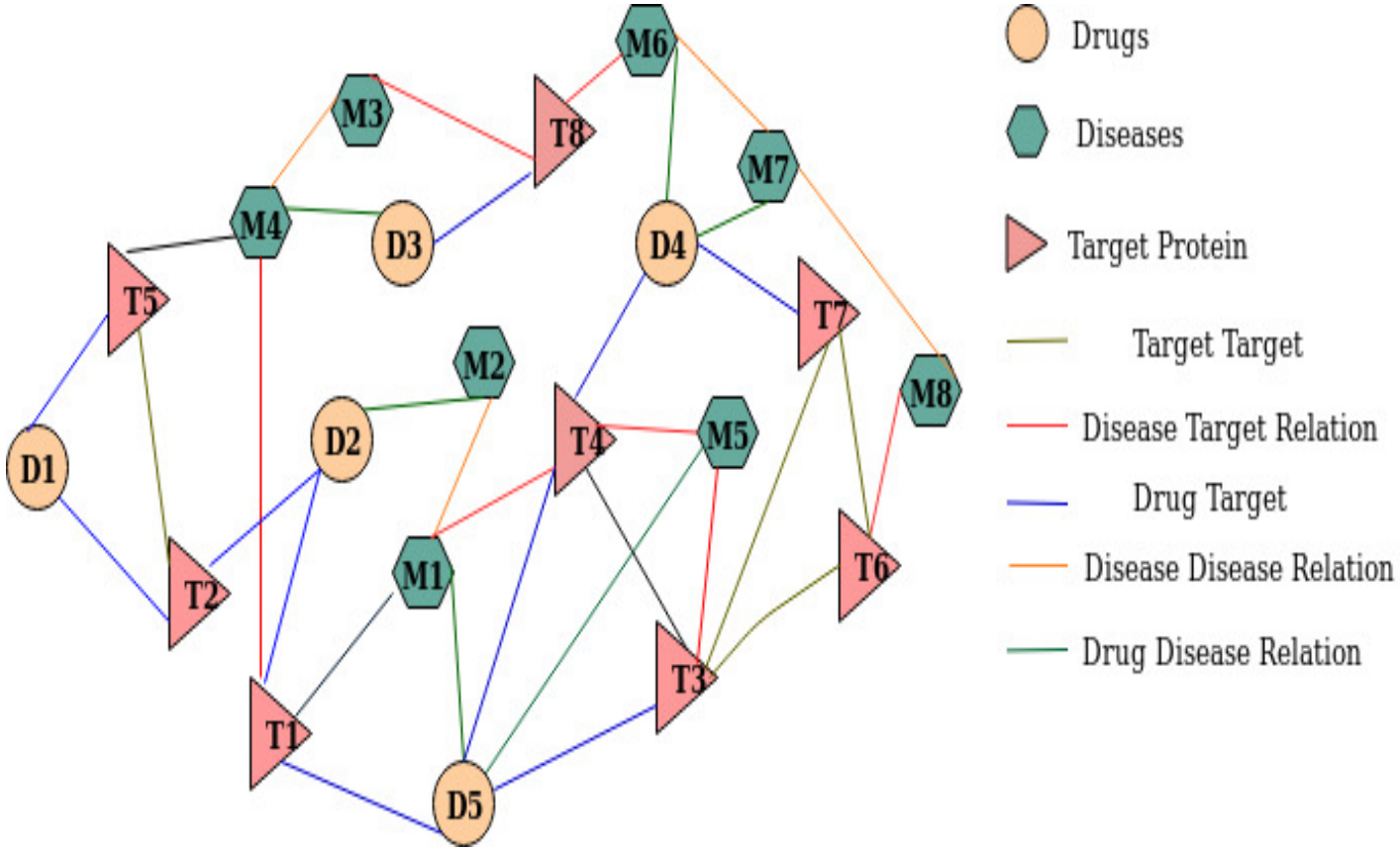


Relation Extraction

Predict the relationship existing between pair of entities **E1** and **E2**



DDI Representation and Prediction using aggregated Heterogeneous Knowledge Graph



Hypothesis

The rich pathway interactions among drugs, targets and diseases are helpful to understand the underlying mechanism of DDI Prediction

Infer higher order relations from the heterogeneous Knowledge Graph comprising of drugs, targets and diseases.

Future of NLP

Transformation from data-driven to intelligence-driven decisions

NLP techcnology will play a key role.

- Fluent human-to-machine interaction for all languages
- Multilingual access
- Harness unstructured data and make it more meaningful to a machine