Ollscoil na Gaillimhe
University of Galway

Advanced Topics in NLP

Introduction

Dr. Paul Buitelaar & Dr. Omnia Zayed

Data Science Institute

University of Galway

University
*of*Galway.ie

Learning Objectives of this Course

- Gain insights into knowledge extraction from text, in particular around entities and relations
- Gain insights into opinion mining, in particular on emotion analysis, dynamic identification of aspect and analysis of figurative language
- Gain insights into language generation, in particular in the context of machine translation and chatbot development

Administrative Issues

Lecturers

Dr. Paul Buitelaar & Dr. Omnia Zayed

Lecture Plan

Date		Lecture
9/1/23		Introduction
16/1/23		Knowledge Extraction I: Entities
23/1/23		Knowledge Extraction II: Relations
30/1/23		Opinion Mining I: Emotions
6/2/23		bank holiday
13/2/23		Opinion Mining II: Aspect
20/2/23		Opinion Mining III: Figurative Language
27/2/23		Language Generation I: Machine Translation I
6/3/23		Language Generation II: Machine Translation II + NLG
13/3/23		Language Generation III: Dialog Systems
20/3/23		Summary
27/3/23		Industry Talk

Labs

Time & Venue: Fridays 11am-1pm, IT102

Practical exercises covering the course content

Prerequisites are basic knowledge of Python, see: https://docs.python.org/3/tutorial/

We will use Google colab Jupyter notebook, see: https://colab.research.google.com

Dhairya Dalal

Ali Hatami

Lab Exercises

Date		Lab
13/1/23		intro to base NLP tools and methods
20/1/23		Named Entity Recognition
27/1/23		relation prediction
3/2/23		emotion classification
10/2/23		no lab - bank holiday
17/2/23		aspect-based sentiment analysis
24/2/23		metaphor classification
3/3/23		build a machine translation model step by step
10/3/23		machine translation evaluation
17/3/23		no lab - bank holiday
24/3/23		chatbot development with RASA

Assignments

Two assignments

- Assignment 1: Released Jan 30th, Due Feb 20th
- Assignment 2: Released Mar 6th, Due Apr 3rd

Assignments count for 50% of final grade

Recommended Reading

Lectures

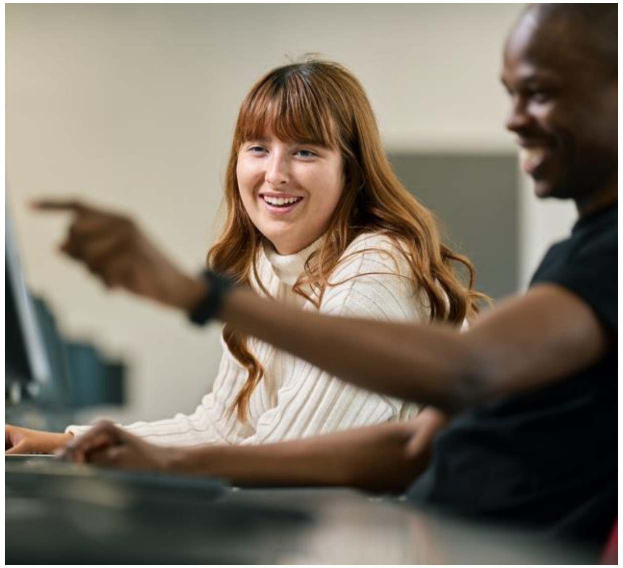
- Jurafsky and Martin, ***SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition***, 3rd edition: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- Chris Manning and Hinrich Schütze, ***Foundations of Statistical Natural Language Processing***, MIT Press. Cambridge, MA: May 1999: https://nlp.stanford.edu/fsnlp/

Labs

- ***Dive into Deep Learning***: https://d2l.ai/

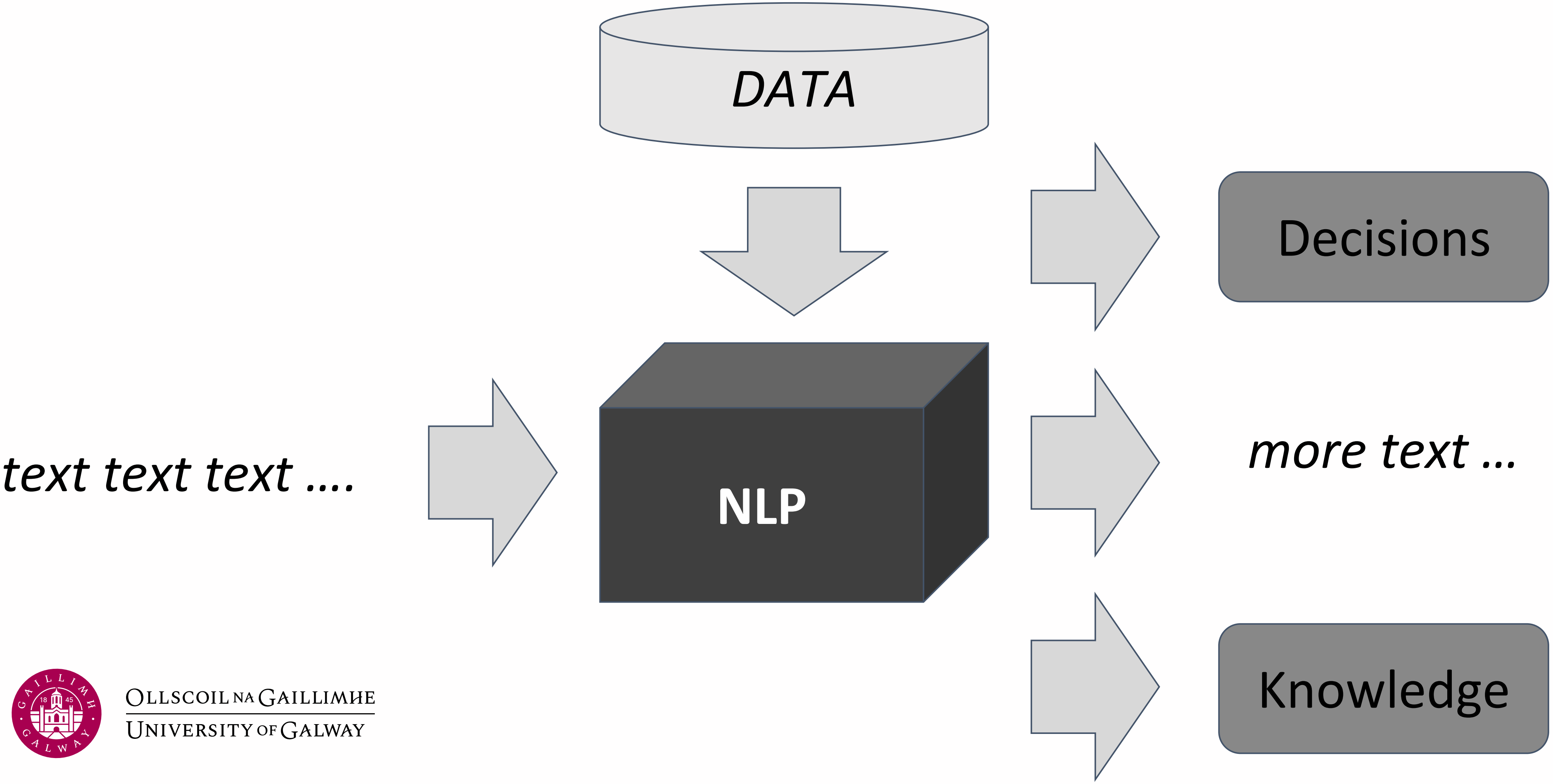
Summary of Intro NLP
“What we learned so far”

Buzz Groups



What have we learned in the Intro NLP course?

Natural Language Processing



Linguistics

text text text ….

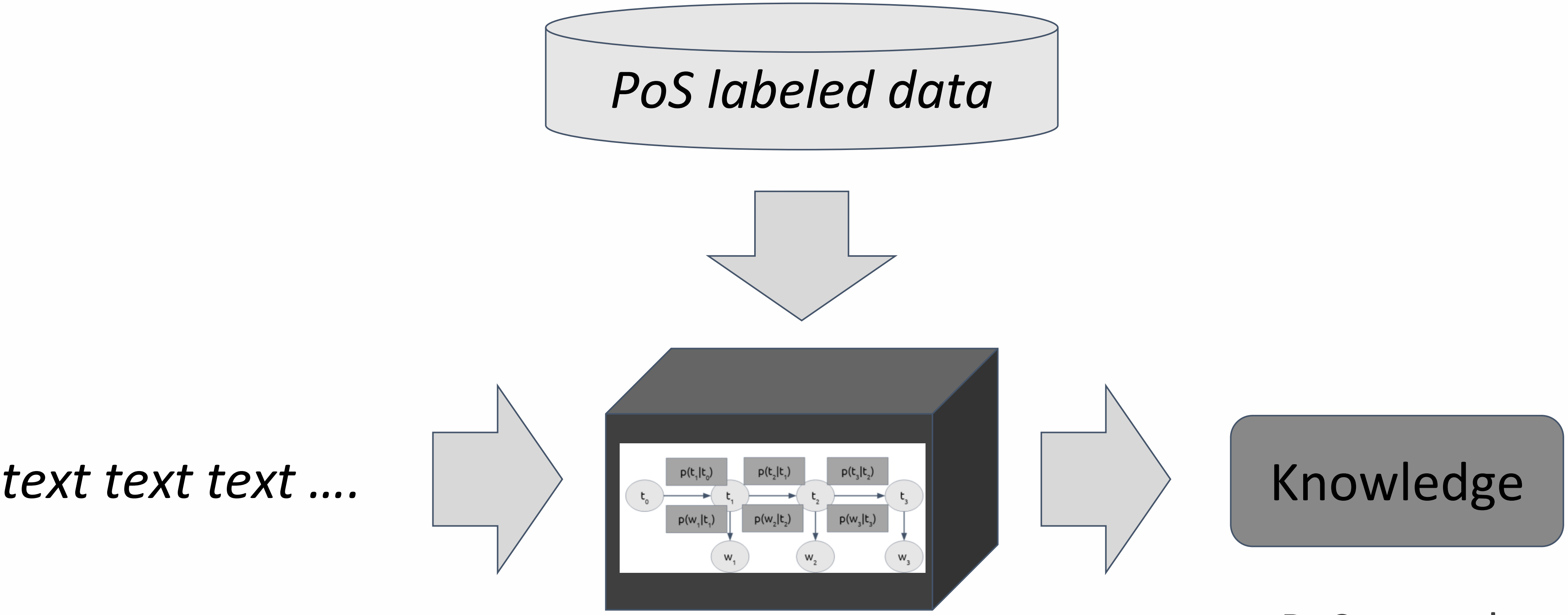
Linguistic Structure & Levels of Analysis

- **Morphology**: tokenization (MWEs), inflection, derivation, stemming, lemmatization
- **Syntax**: part-of-speech, grammar (constituency vs. dependency)
- **Semantics**: word sense, semantic roles, coreference

Language Data

- **Lexicon**: WordNet, FrameNet
- **Corpora**: annotation (data labeling), multilingual, domain-specific

Syntax: Part of Speech



PoS labeled data

text text text ....

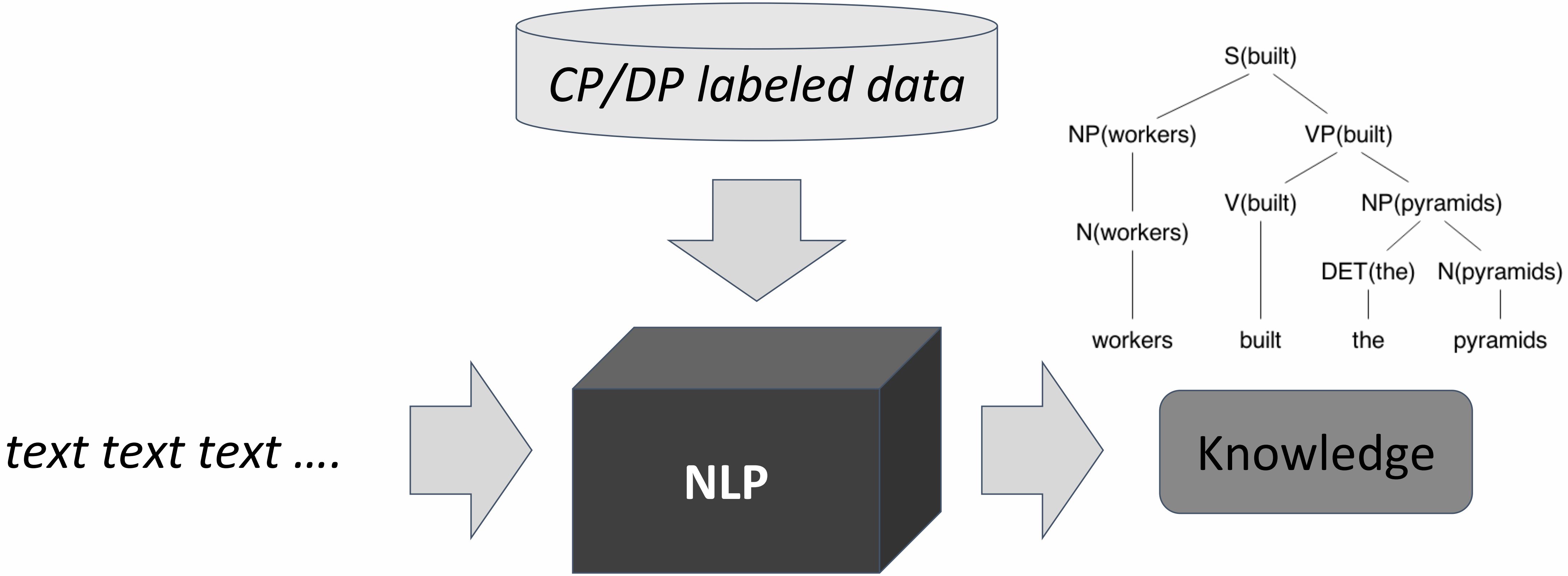
Knowledge

- Hidden Markov Model

- PoS tagged text

Ollscoil na Gaillimhe
University of Galway

Syntax: Constituency / Dependency Parsing



CP/DP labeled data

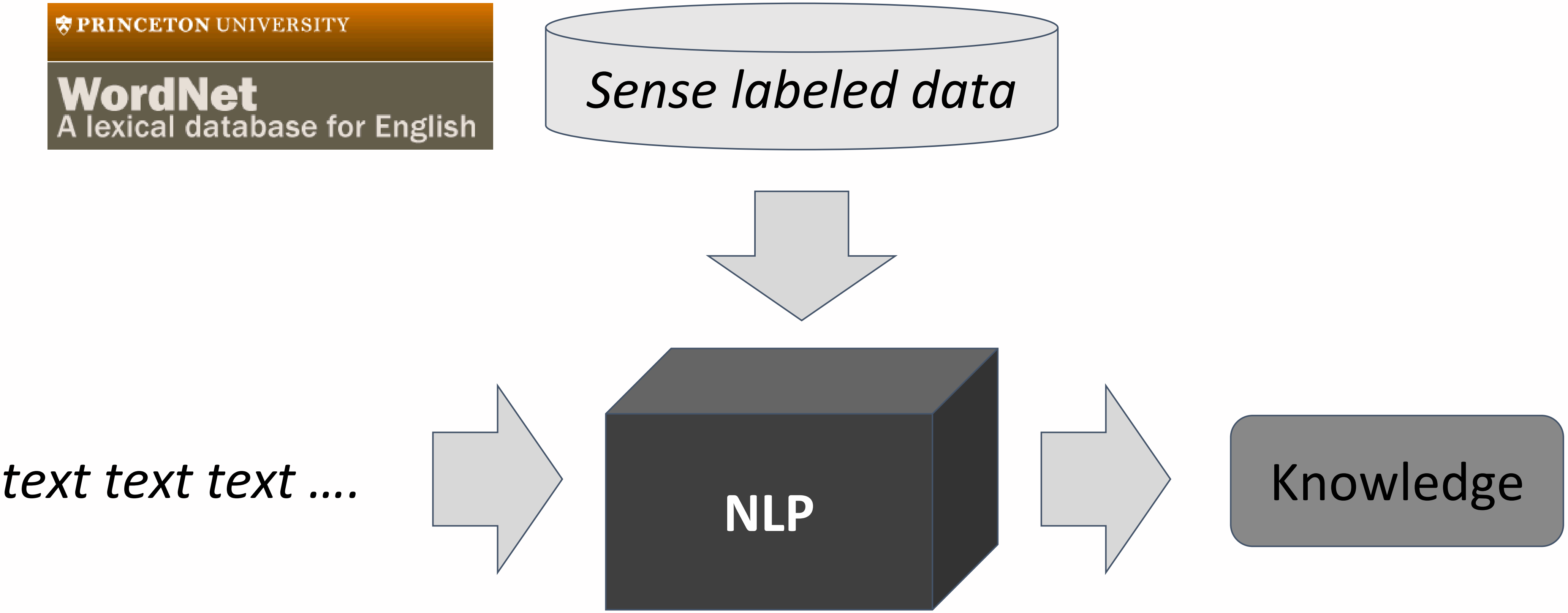
text text text ....

NLP

Knowledge

- CYK algorithm
- text annotated with parse trees

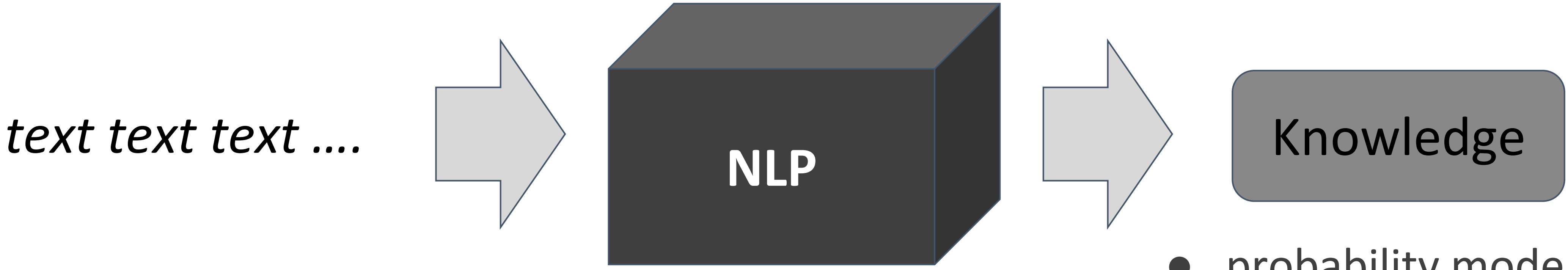
Semantics: Word Sense Disambiguation



- word sense annotated text

- Lesk algorithm

Probability



text text text ….

NLP

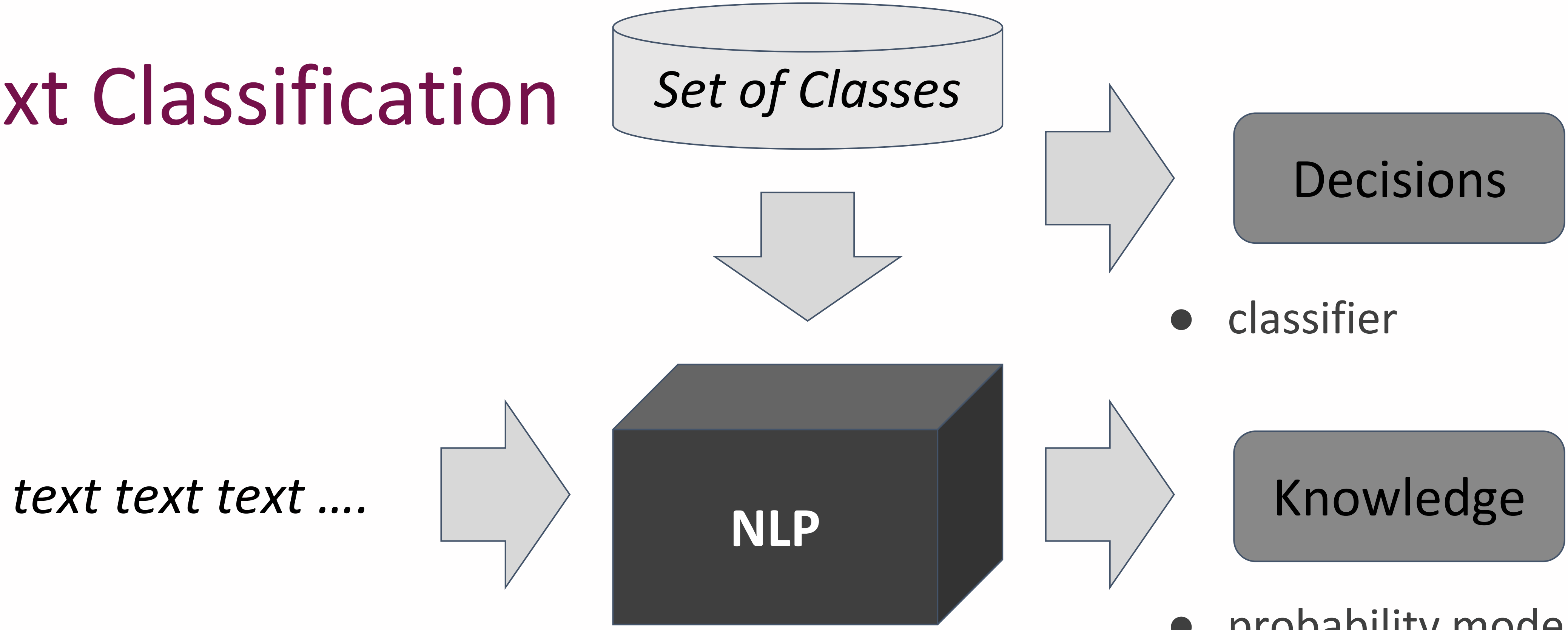
Knowledge

- probability model

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)} \qquad TF\text{-}IDF = f_w \times \left(\log\left(\frac{N}{N_w}\right) + 1\right)$$

- estimate word probability using Naive Bayes, TF-IDF

Text Classification



Set of Classes

text text text ….

NLP

Decisions

- classifier

Knowledge

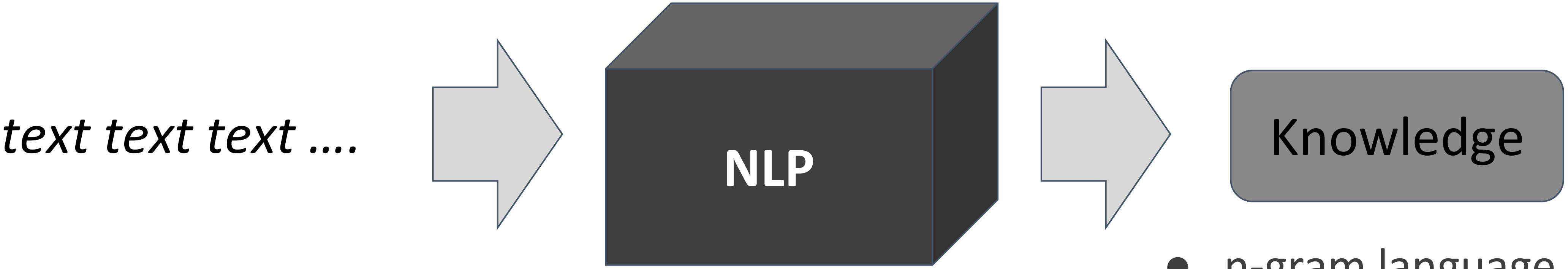
- probability model

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)}$$

$$TF\text{-}IDF = f_w \times \left(\log\left(\frac{N}{N_w}\right) + 1\right)$$

- estimate word probability using Naive Bayes, TF-IDF

Language Modeling



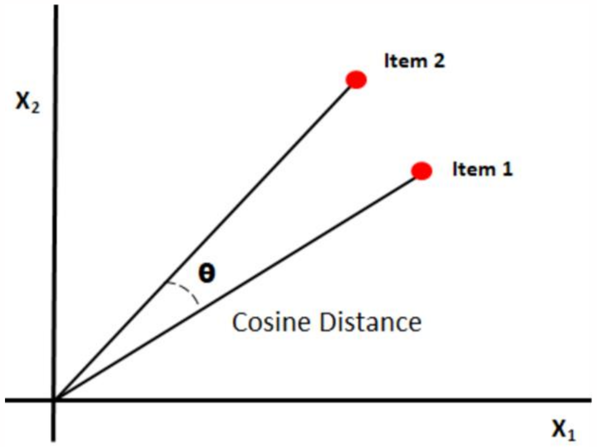
- n-gram language model

$$p(w_1 w_2 \ldots w_n) = \prod_{k=1,\ldots,n} p(w_k | w_{k-m+1} \ldots w_{k-1})$$

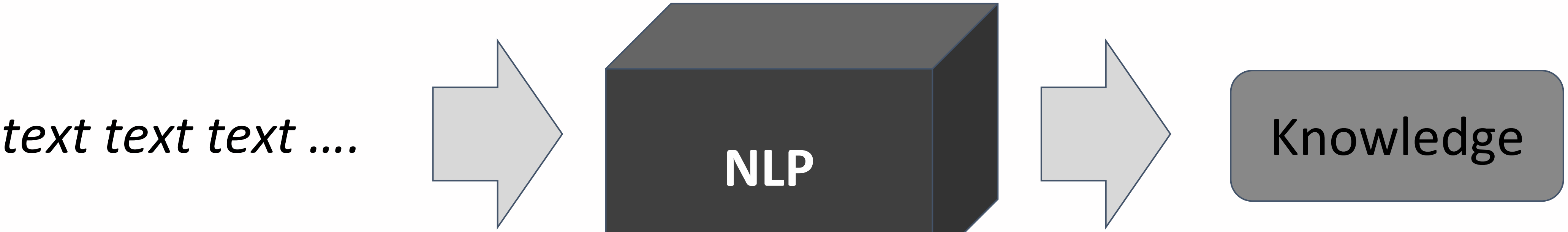
- estimate the probability of a sentence using n-gram model

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Vector Space



- Cosine similarity


text text text ….

NLP

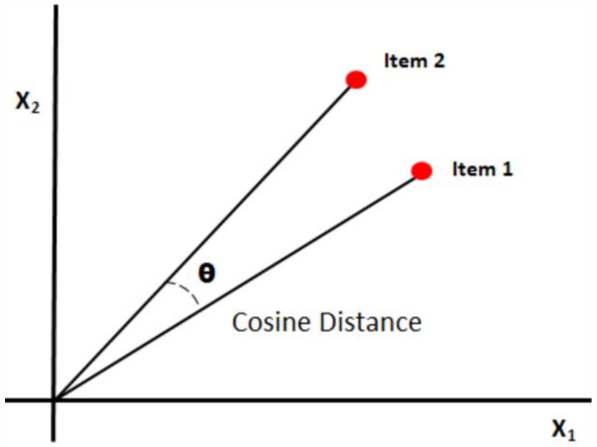
Knowledge

- distributional model

- co-occurrence matrix

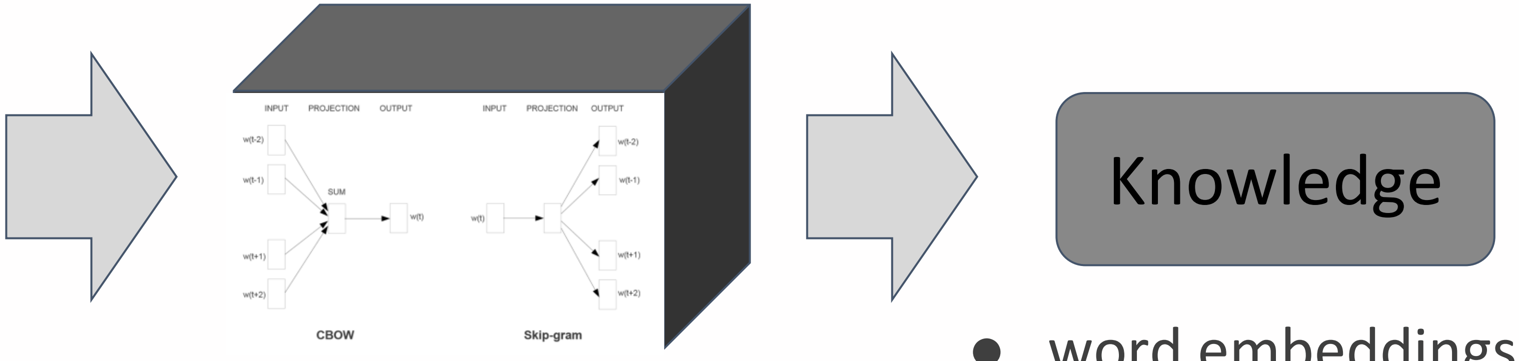
Image from https://www.oreilly.com/library/view/statistics-for-machine

Word Embeddings



- Cosine similarity

text text text ....



Knowledge

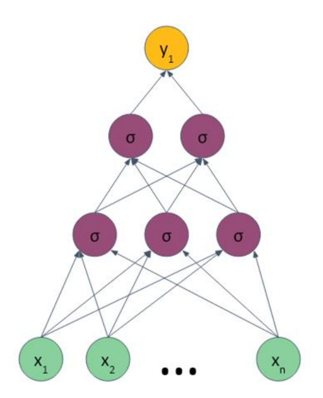
- word embeddings

- neural network architecture

Image from https://www.oreilly.com/library/view/statistics-for-machine

RNNs: Text Classification

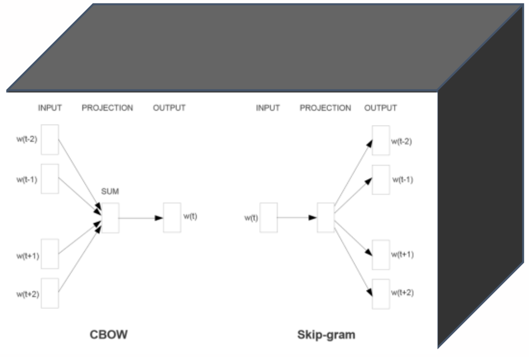
Set of Classes



Decisions

- RNN based classifier

text text text ....



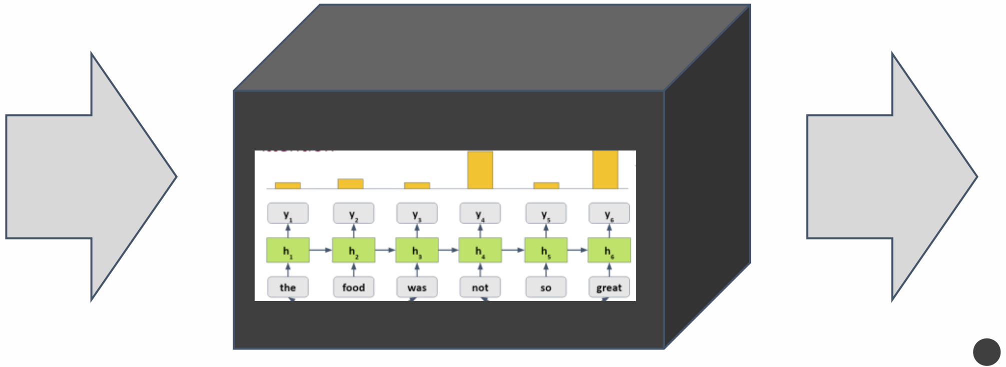
Knowledge

- word embeddings

- neural network architecture

Deep Learning: Transformers

text text text ....



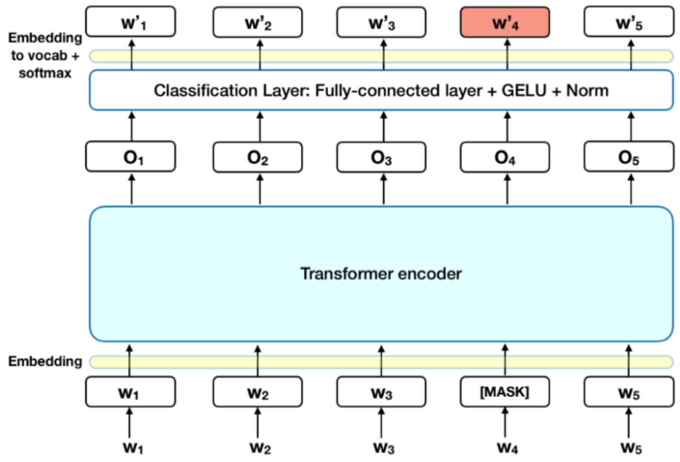
Knowledge

- Pretrained model

- Transformer with self-attention

Transformers: Text Classification

Set of Classes

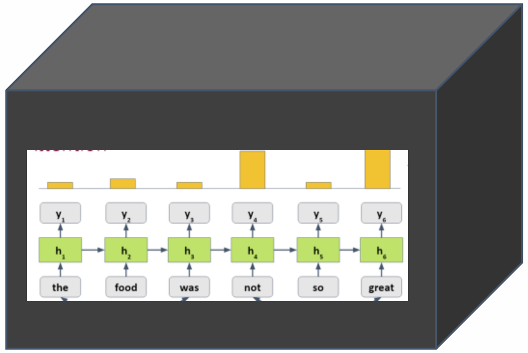


w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

Decisions

- Classifier with fine-tuning for transfer learning

text text text ….



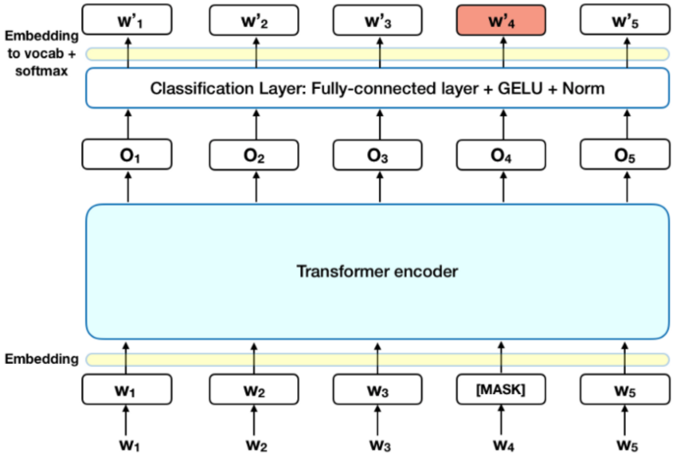
Knowledge

- Pretrained model

- Transformer with self-attention

Applications: Sentiment Analysis

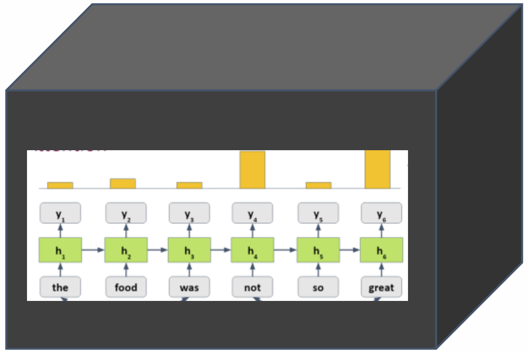
SA labeled data


w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

Decisions

- Sentiment analysis classifier

text text text ….

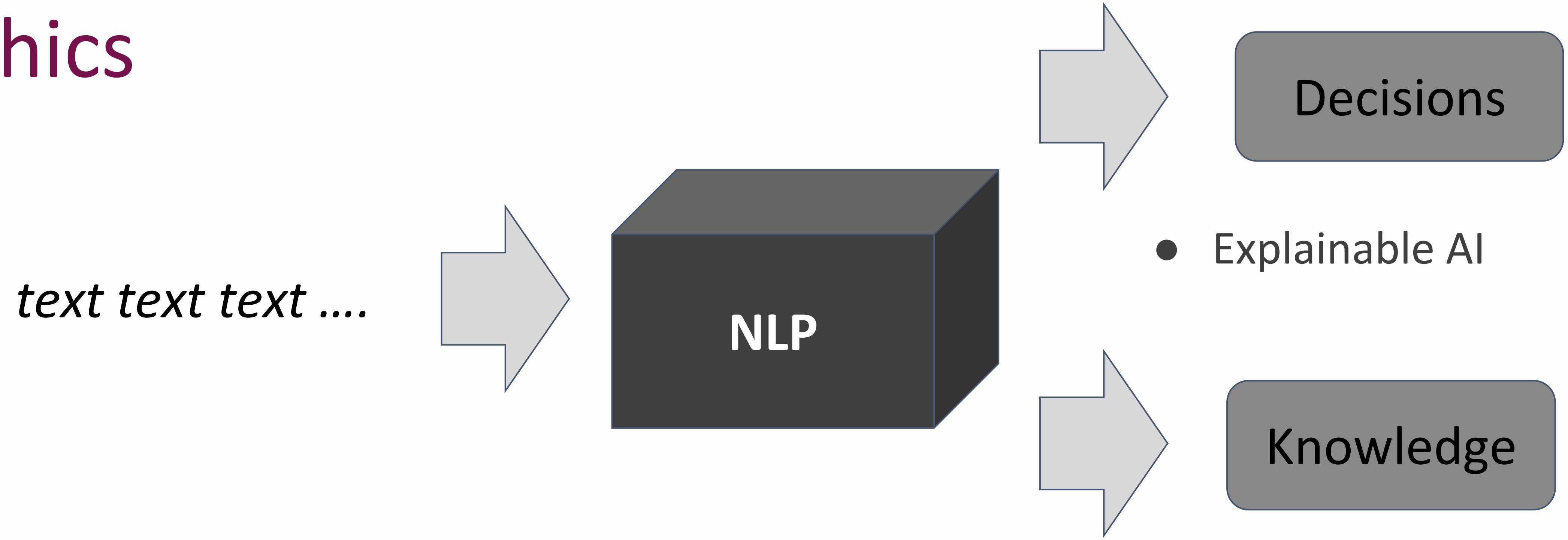


Knowledge

- Pretrained model

- Transformer with self-attention

Ethics



text text text ....

NLP

Decisions

Knowledge

- Explainable AI

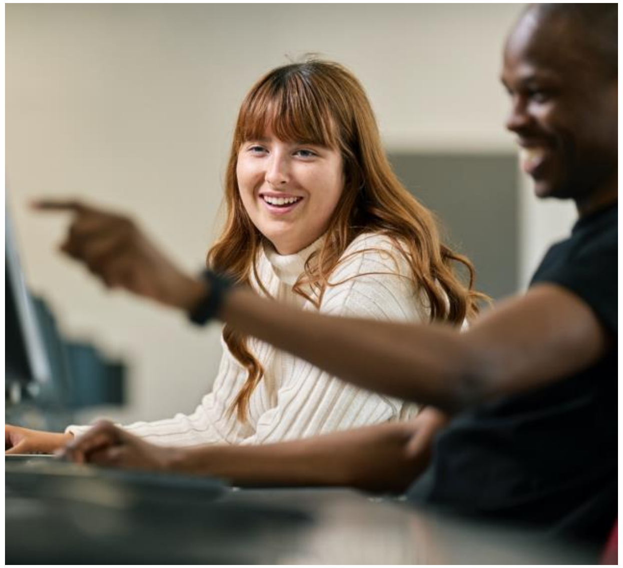
- Data privacy (GDPR)
- Data Protection Impact Assessment
- Data Statement

- Ethical, bias-aware NLP

- Trustworthy AI

Ollscoil na Gaillimhe
University of Galway

Buzz Groups



What is missing?

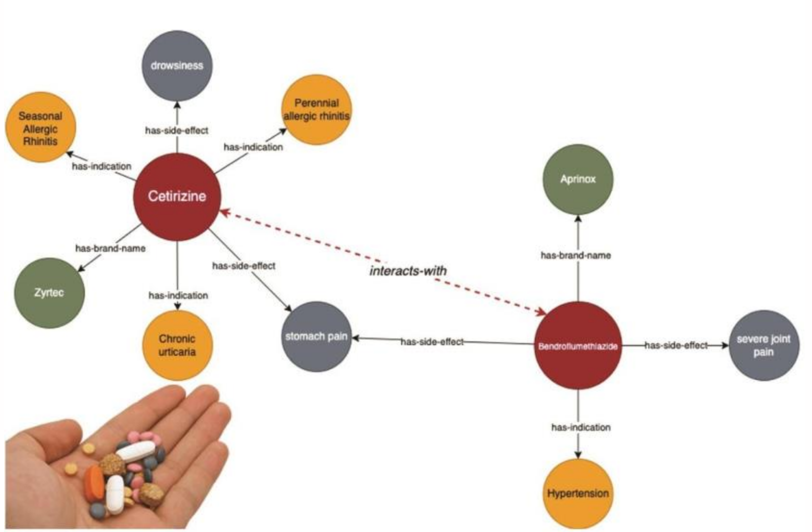
NLP tasks missing

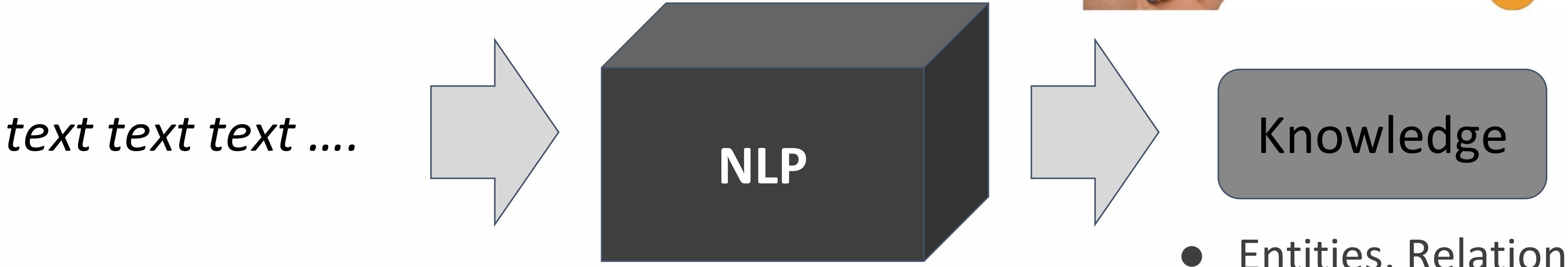
“Focus of this course”

Knowledge Extraction Tasks

Symbolic Knowledge

- Named Entity Recognition & Entity Linking
- Relation Extraction





- Entities, Relations

Image source: https://timesofmalta.com/articles/view/identifying-drug-interactions-using-knowledge-graphs.936927

Knowledge Extraction - Definition

Creation of knowledge from **unstructured (textual, language) data**

Extracted knowledge must **facilitate inferencing**

Requires **reuse and/or generation of formal knowledge**

Adapted from *https://en.wikipedia.org/wiki/Knowledge_extraction*

Knowledge Extraction - Example

“Naomi Carey, who is the director of Hutchinson Care Homes said she is currently only able to operate at 85% capacity.”

“Galway GP Martin Daley, former president of the Irish Medical Organisation, gave a statement today to this effect.”

Entities	PERSON:	*Martin Daley, Naomi Carey*
	ORGANISATION:	*Irish Medical Organisation, Hutchinson Care Homes*
	CITY:	*Galway*

Relations

Martin Daley → at-organisation ← *Irish Medical Organisation*

Naomi Carey → at-organisation ← *Hutchinson Care Homes*

Martin Daley → has-occupation ← *GP*

Naomi Carey → has-occupation ← *former president*

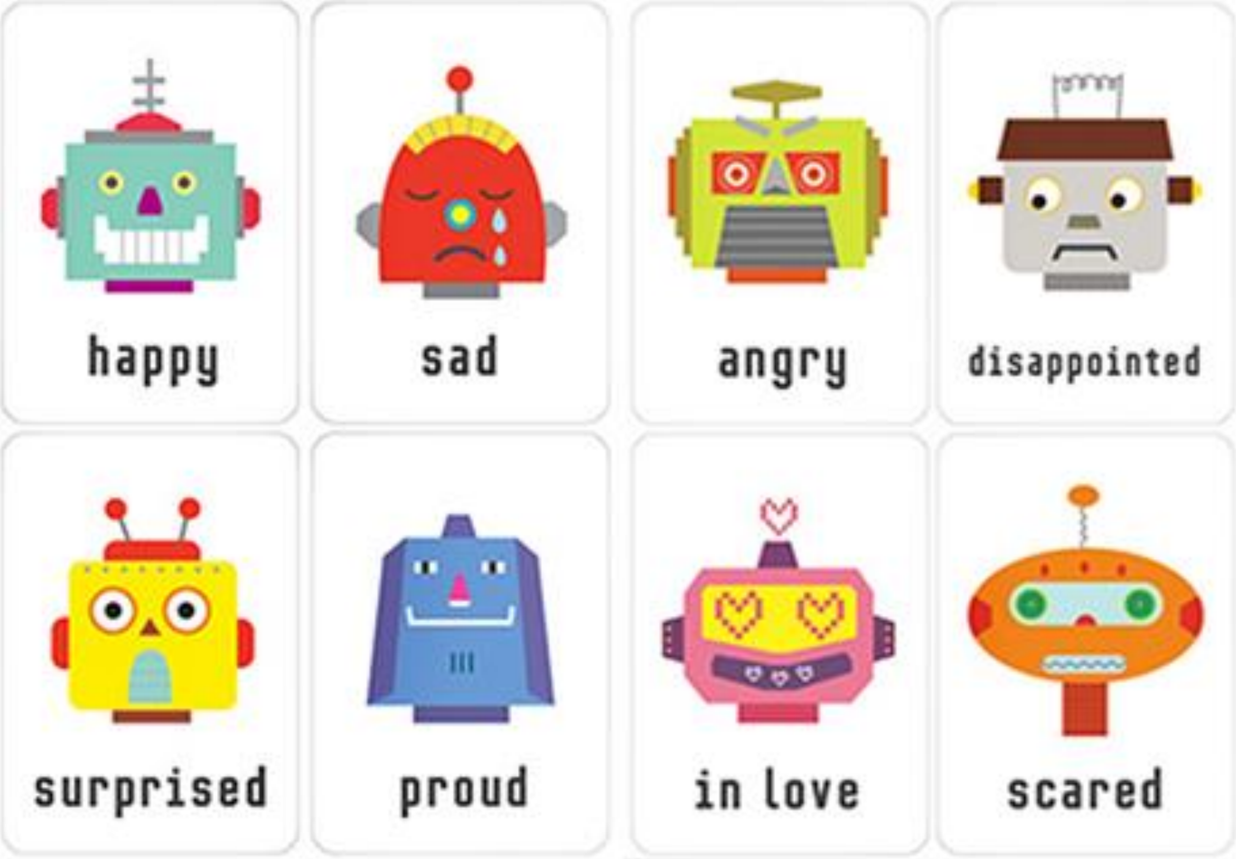
Opinion Mining Tasks

Emotion Analysis

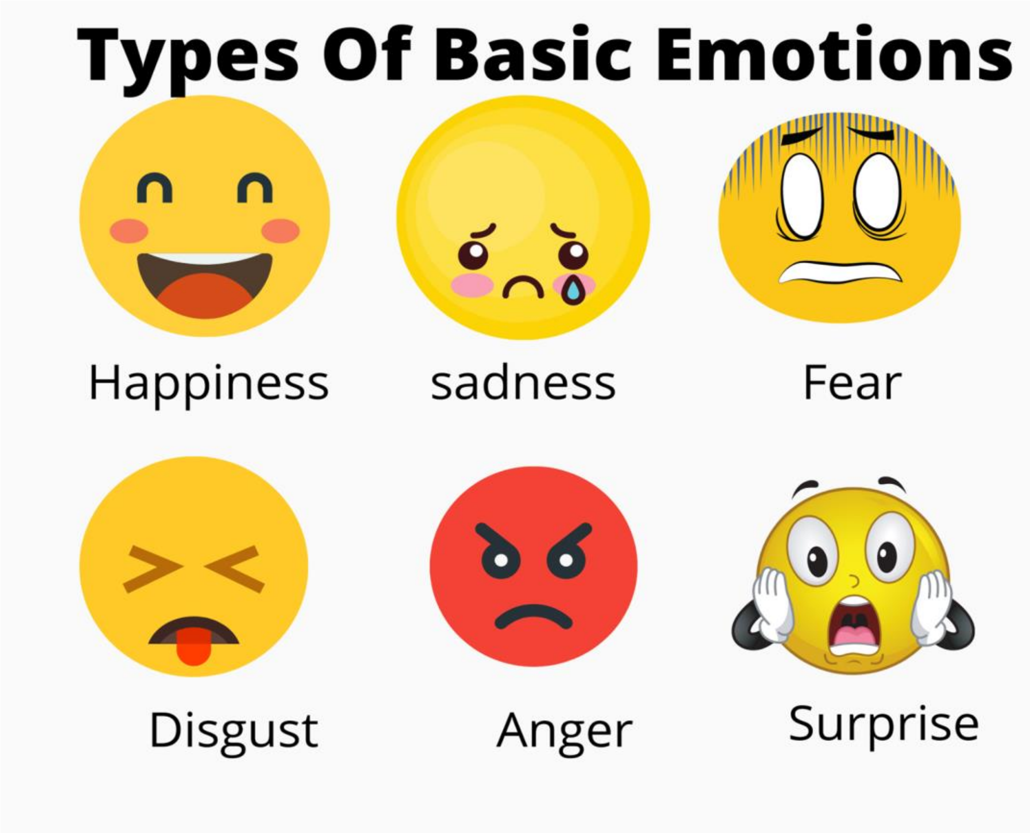
Aspect Mining

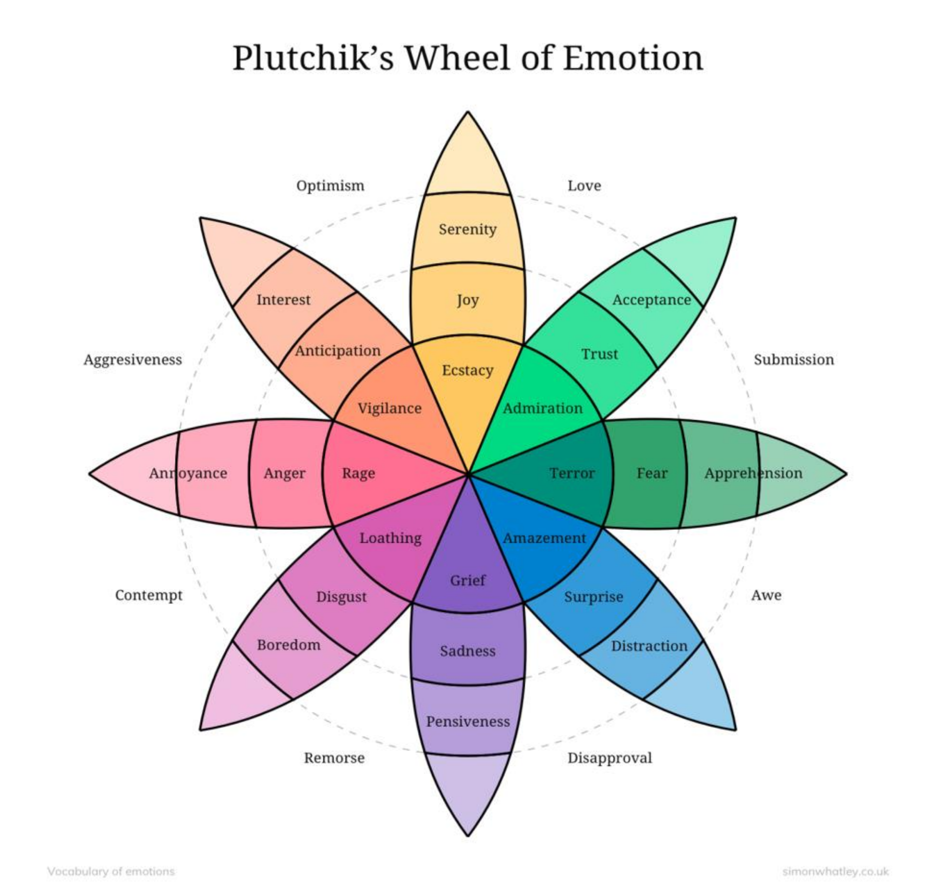
Figurative Language Processing

Emotion Analysis

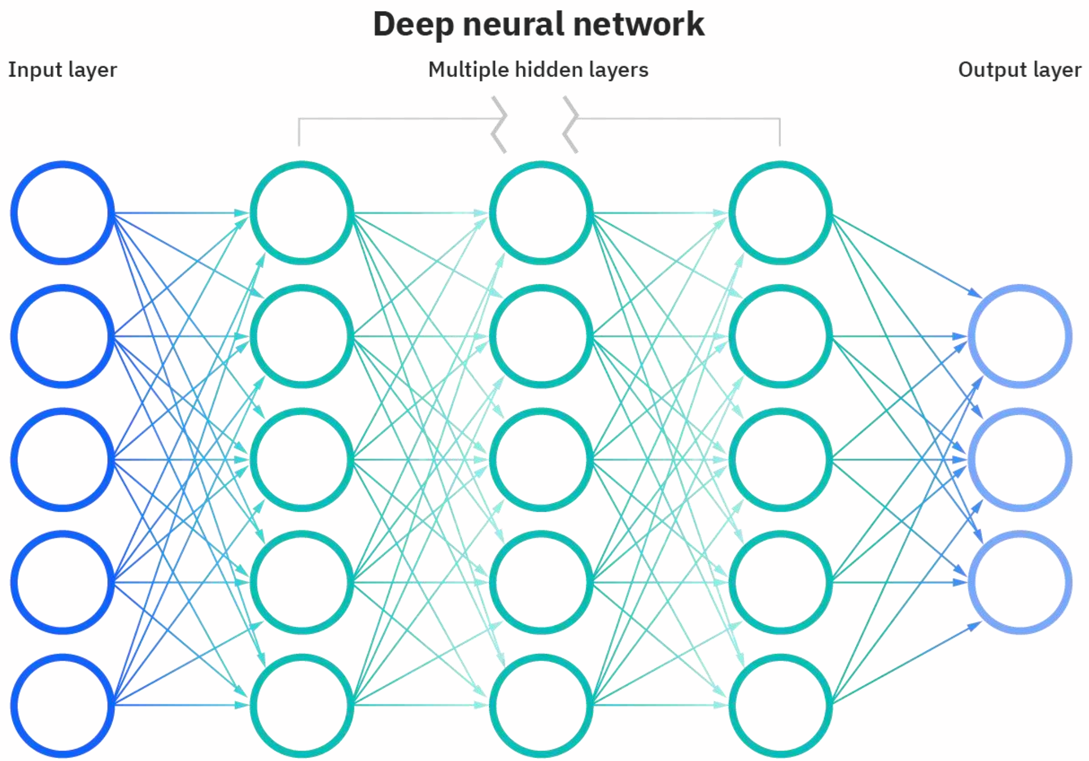


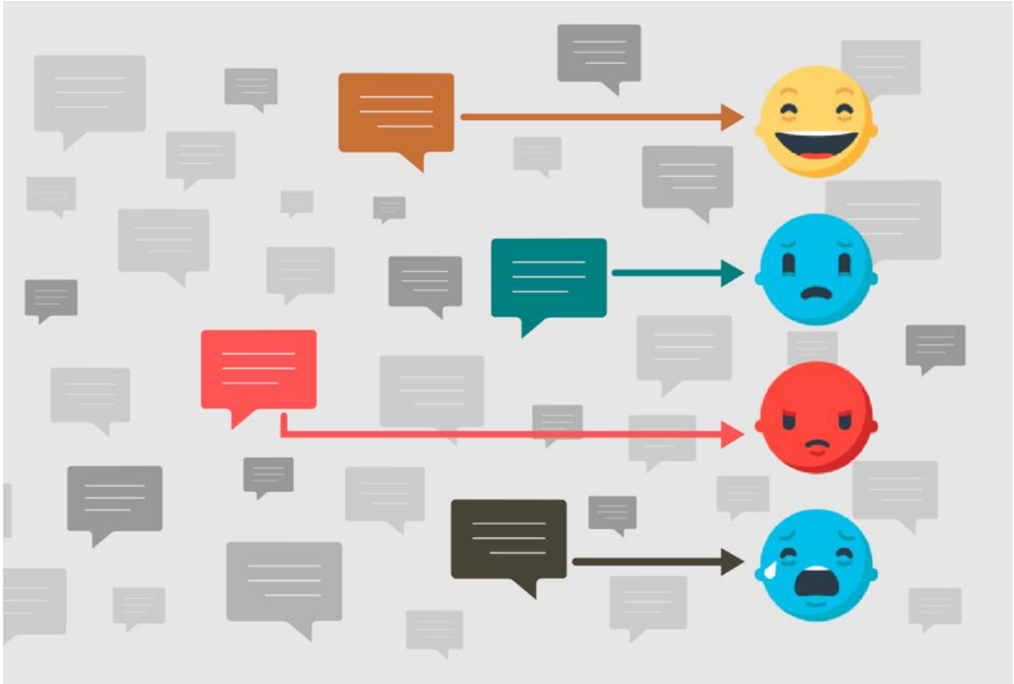
Emotion Analysis





Emotion Analysis



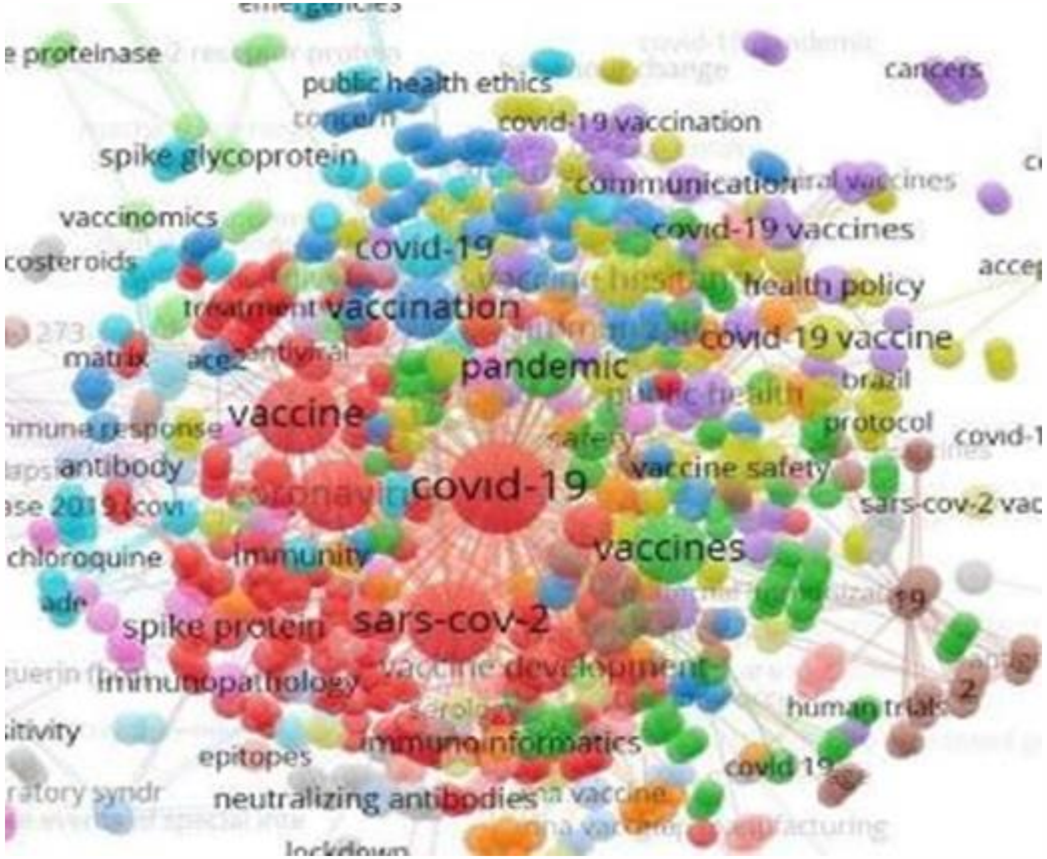


OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

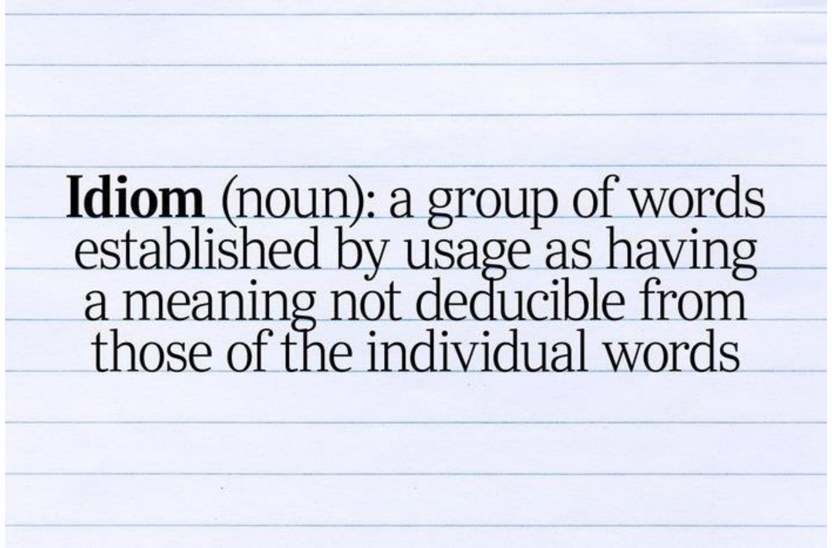
Aspect Mining

*“The camera’s focus was **bad**, but has a **great** size and is **easy-to-use**.”*

Aspect Mining

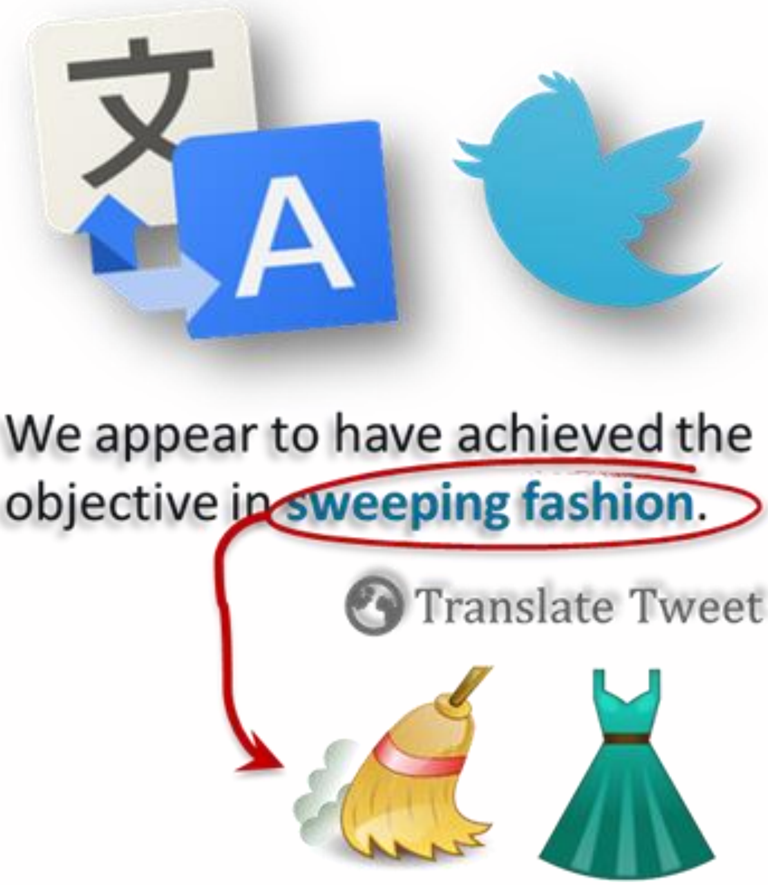


Figurative Language Processing



Idiom (noun): a group of words established by usage as having a meaning not deducible from those of the individual words

Figurative Language Processing



We appear to have achieved the objective in sweeping fashion.

Language Generation Tasks

Machine Translation

Data-to-Text Generation

Dialog System Development

Language Generation: Machine Translation

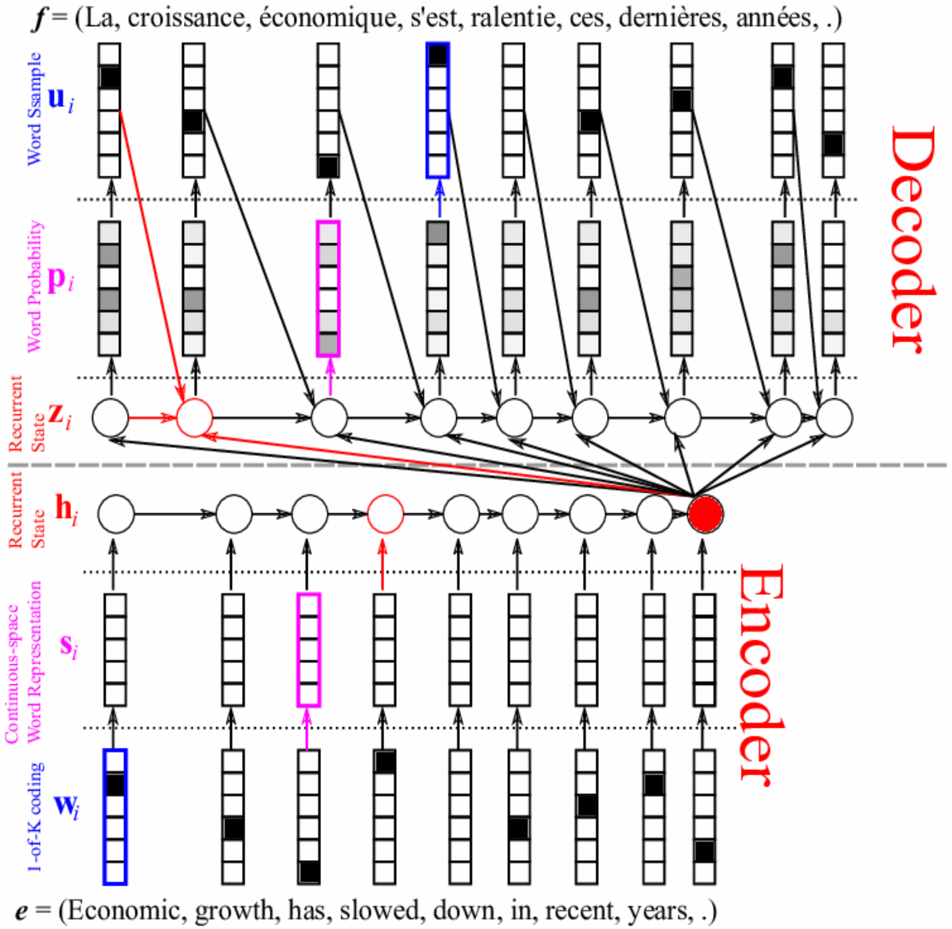
English text ….

NLP

Nederlandse tekst …

Machine Translation

```
gerade zu diesem Stamm gehören ||| belong just to these families ||| 0.390442 2.9025
gerade zu diesem Stamm gehören ||| belong just to these ||| 0.390442 2.9025e-15 0.2
gerade zu diesem Stamm gehören ||| them belong just to these ||| 0.390442 2.9025e-
gerade zu diesem Stamm ||| belong just to these families ||| 0.390442 1.5563e-11 0.2
gerade zu diesem Stamm ||| belong just to these ||| 0.390442 1.5563e-11 0.260295 9.
gerade zu diesem Stamm ||| them belong just to these ||| 0.390442 1.5563e-11 0.260
gerade zu diesem ||| belong just to ||| 0.390442 1.33531e-08 0.260295 4.9937e-09 2.
gerade zu diesem ||| of them belong just to ||| 0.390442 1.33531e-08 0.260295 4.79
gerade zu diesem ||| them belong just to ||| 0.390442 1.33531e-08 0.260295 4.6096
gerade zu ||| belong just to ||| 0.390442 1.8515e-05 0.260295 4.9937e-09 2.718 |||
gerade zu ||| of them belong just to ||| 0.390442 1.8515e-05 0.260295 4.79751e-13
gerade zu ||| them belong just to ||| 0.390442 1.8515e-05 0.260295 4.60968e-12 2.7
```

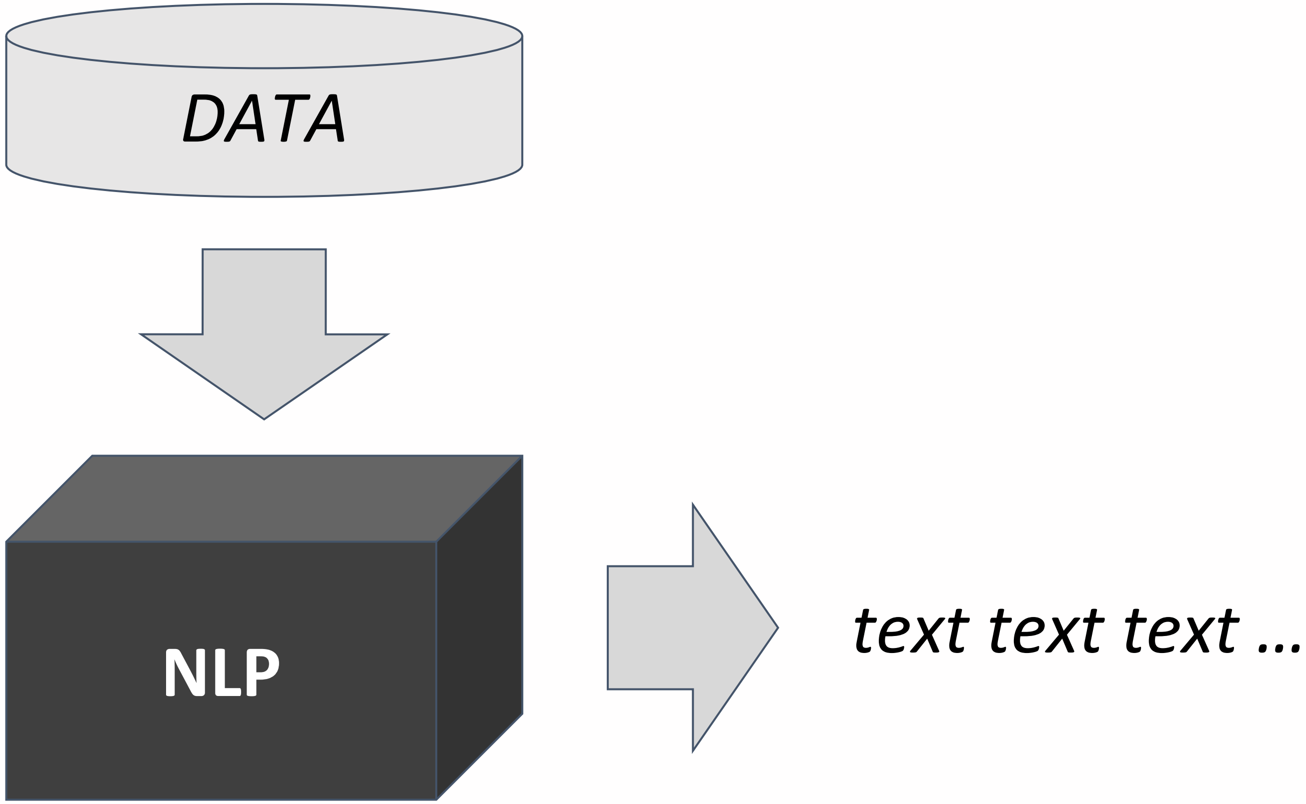


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
$\mathbf{u}_i$
Word Probability
$\mathbf{p}_i$
Recurrent State
$\mathbf{z}_i$
Decoder
Recurrent State
$\mathbf{h}_i$
Continuous-space Word Representation
$\mathbf{s}_i$
1-of-K coding
$\mathbf{w}_i$
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Statistical Machine Translation

Neural Machine Translation

Language Generation: Data-to-Text



Data-to-Text Generation

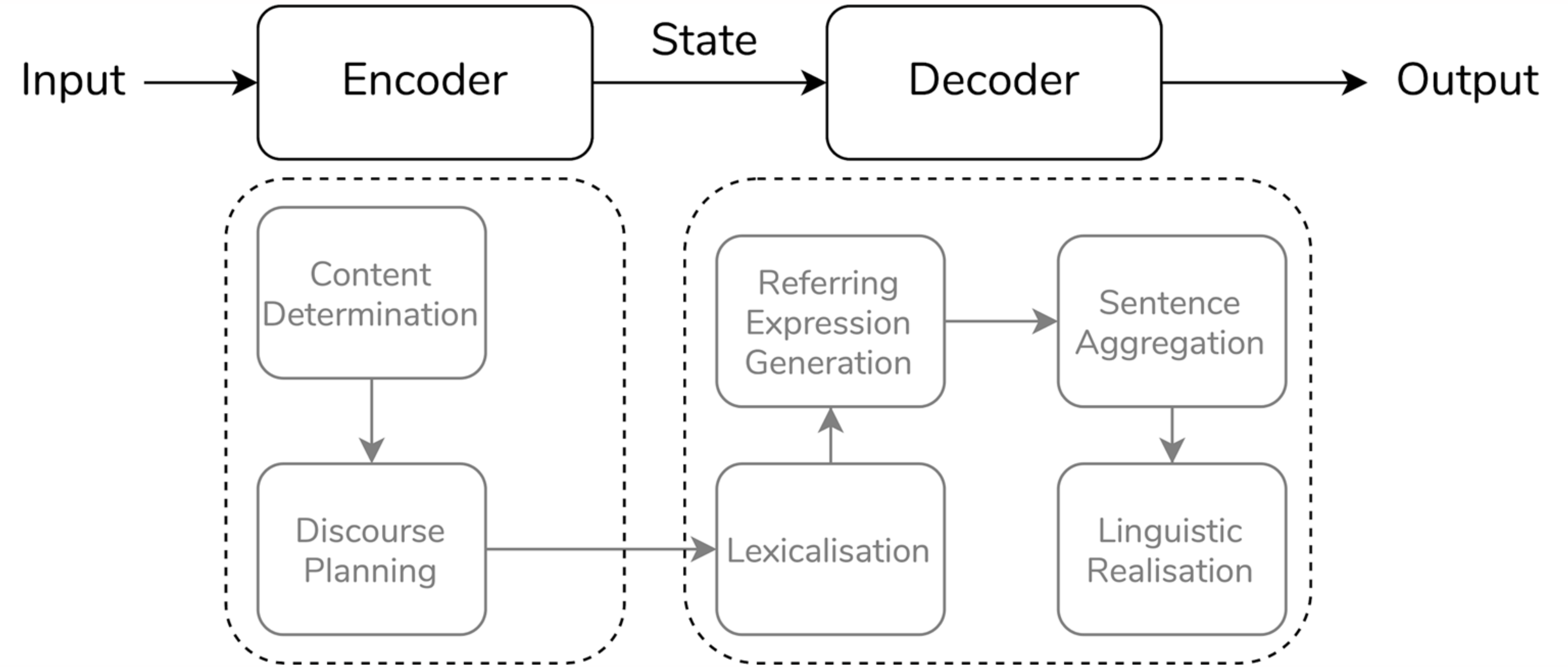
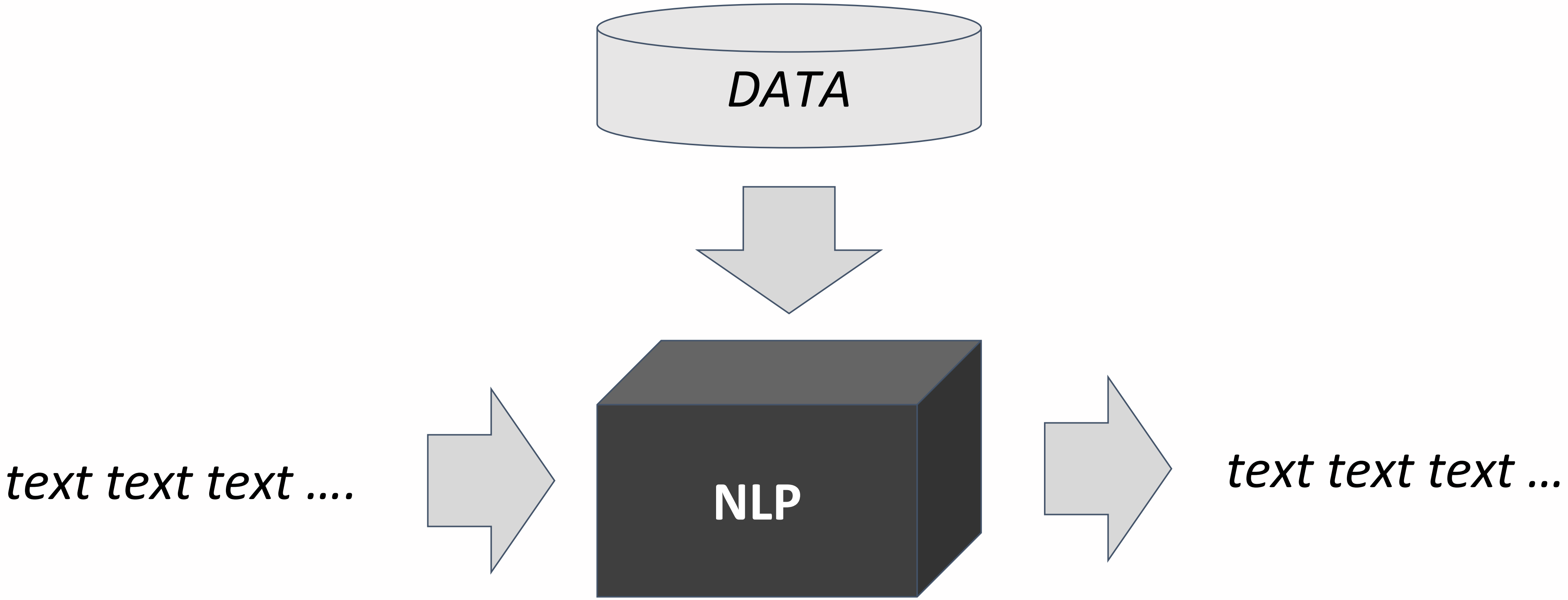


Image by Nivranshu Pasricha

Language Generation: Dialog Systems



DATA

text text text ….

NLP

text text text …

Dialog System Development

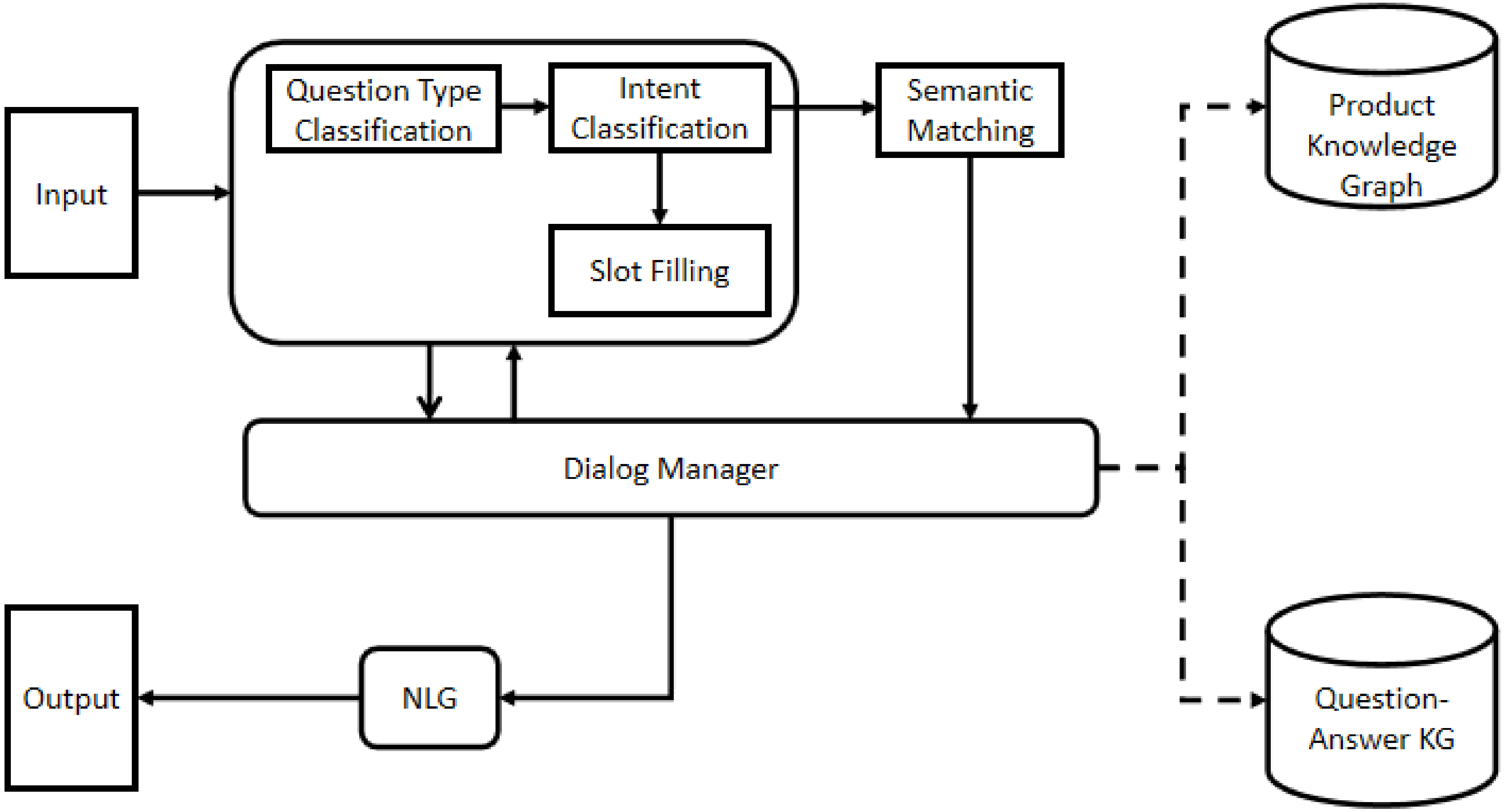
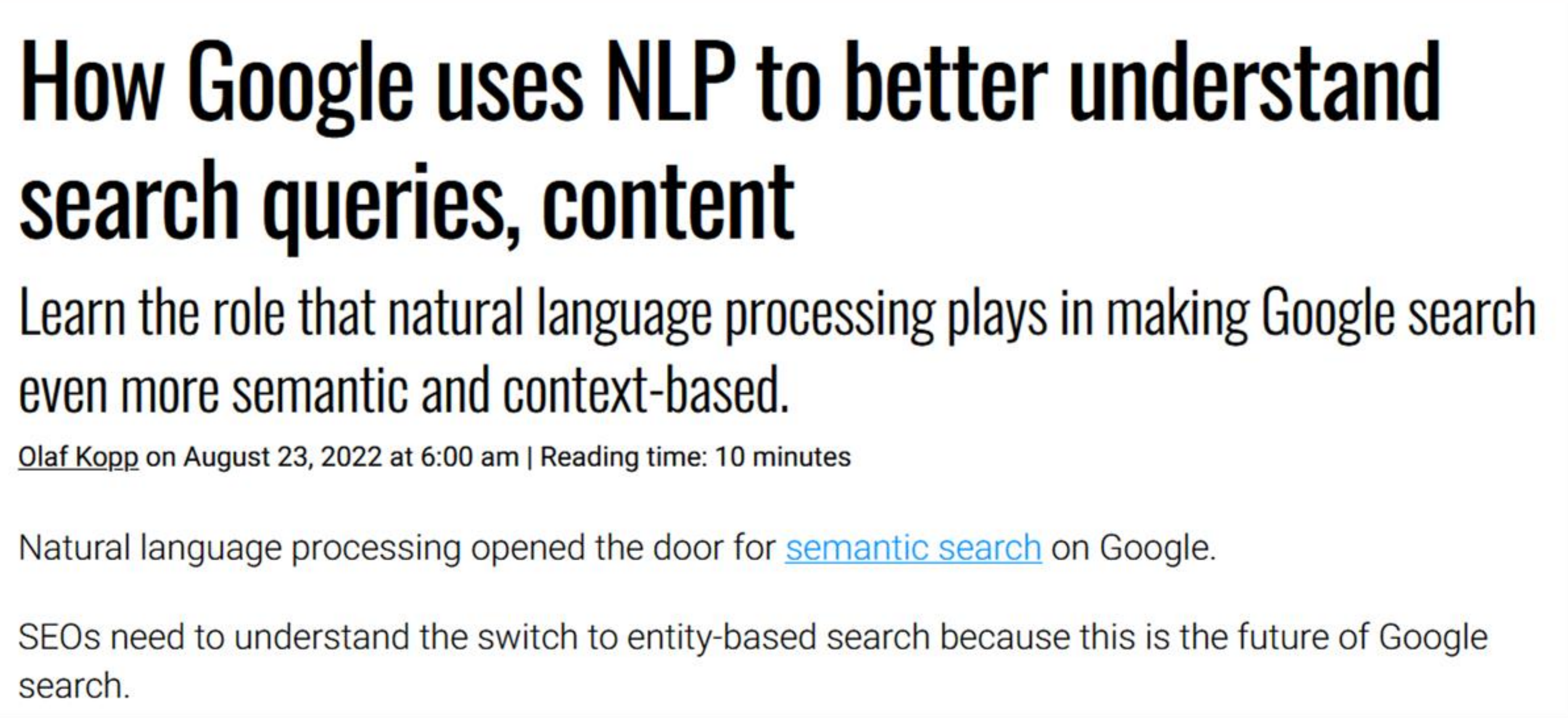


Image from: Zhao, G., Zhao, J., Li, Y., Alt, C., Schwarzenberg, R., Hennig, L., Schaffer, S., Schmeier, S., Hu, C. and Xu, F., 2019. MOLI: Smart Conversation Agent for Mobile Customer Service. *Information*, 10(2), p.63.

NLP in industry



How Google uses NLP to better understand search queries, content

Learn the role that natural language processing plays in making Google search even more semantic and context-based.

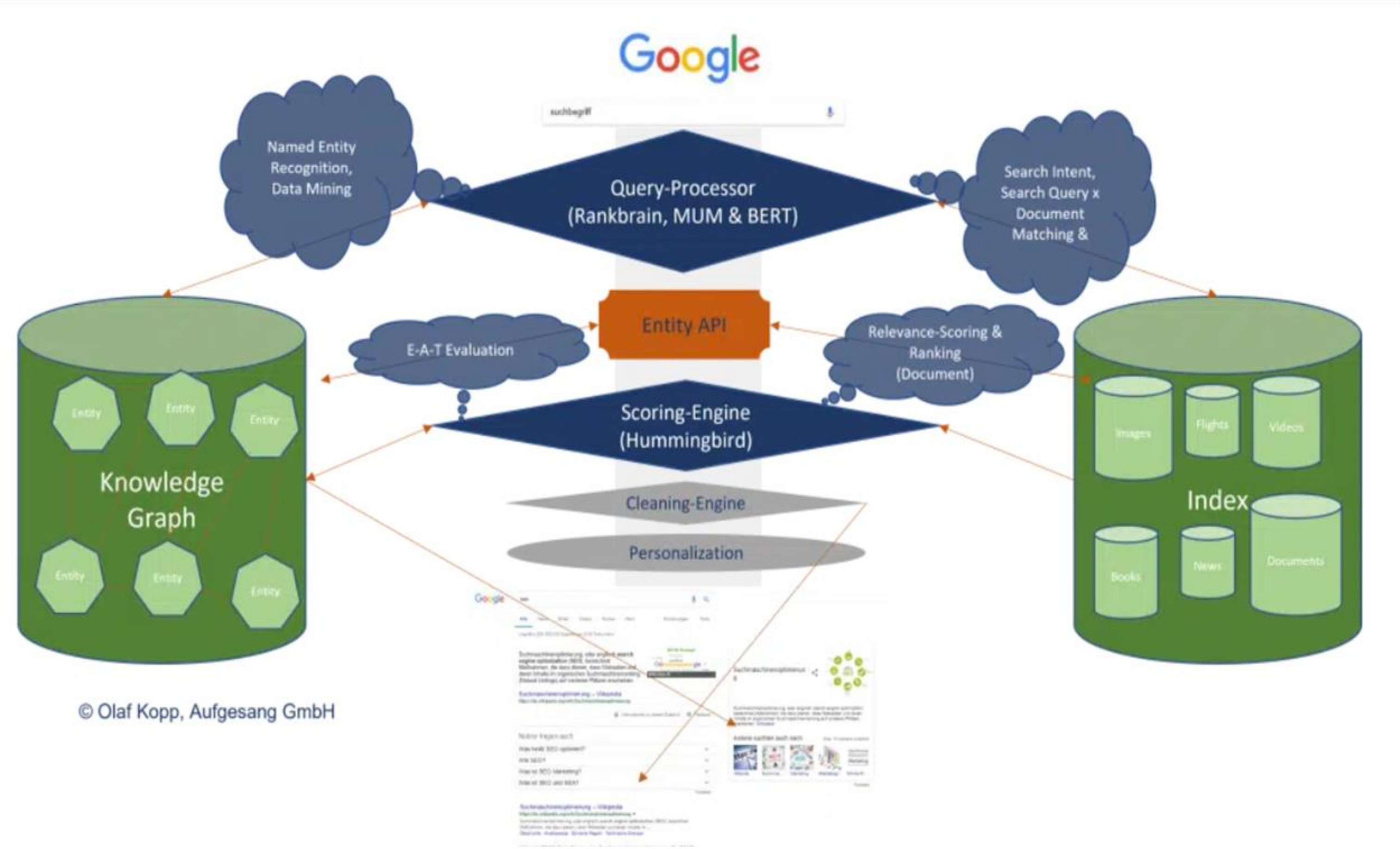
Olaf Kopp on August 23, 2022 at 6:00 am | Reading time: 10 minutes

Natural language processing opened the door for semantic search on Google.

SEOs need to understand the switch to entity-based search because this is the future of Google search.

https://searchengineland.com/how-google-uses-nlp-to-better-understand-search-queries-content-387340

Google Index & Knowledge Graph


Google
suchbegriff
Named Entity Recognition, Data Mining
Query-Processor (Rankbrain, MUM & BERT)
Search Intent, Search Query x Document Matching &
E-A-T Evaluation
Entity API
Relevance-Scoring & Ranking (Document)
Scoring-Engine (Hummingbird)
Cleaning-Engine
Personalization
Entity
Entity
Entity
Knowledge Graph
Entity
Entity
Entity
Images
Flights
Videos
Index
Books
News
Documents

© Olaf Kopp, Aufgesang GmbH

https://searchengineland.com/how-google-uses-nlp-to-better-understand-search-queries-content-387340

“Use of NLP in Google Search”

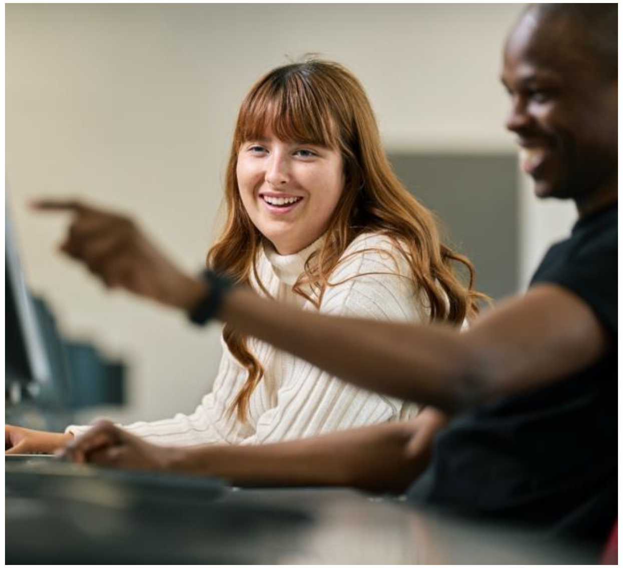
According to Olaf Kopp of Aufgesang GmbH (article in Search Engine Land), Google Search uses NLP for the following:

- Interpretation of search queries.
- Classification of subject and purpose of documents.
- Entity analysis in documents, search queries and social media posts.
- Generating featured snippets and answers in voice search.
- Interpretation of video and audio content.
- Expansion and improvement of the Knowledge Graph.

https://searchengineland.com/how-google-uses-nlp-to-better-understand-search-queries-content-387340

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Buzz Groups



Which NLP tasks are needed for each of these steps?

Interpretation of search queries

“calories in bar of chocolat”

- Relation Extraction, Parsing, Part-of-Speech, WSD

“Biden”

- Entity Linking

“great movies” “scary movie”

- Sentiment / Emotion Analysis

Classification of subject, purpose of documents

legal, health, … documents

- Text Classification, Probability, Language Modeling, Word Embeddings

documents about concept XYZ

- Concept Extraction

Entity analysis in documents, queries and posts

identify names of people, locations, things

- NER, Entity Linking, Concept/Taxonomy Extraction

identify relations between people, locations, things

- Relation Extraction, Parsing, Semantic Role Labeling

resolution of pronouns

- Coreference Resolution

Generating featured snippets and answers

summarize one or more retrieved documents

- Text Summarization

generate a specific answer

- Natural Language Generation, Dialog

translate a text

- Machine Translation

Interpretation of video and audio content

retrieve relevant videos for a search query

- Multimodal Analysis

speech interaction

- Speech-to-Text, Dialog

Expansion/improvement of Knowledge Graph

include new entities

- Entity Linking, Concept Extraction, Taxonomy Extraction/Extension

include new or update existing relations

- Relation Extraction, Knowledge Graph Completion

Lab of this week

Intro to base NLP tools and methods

Tools:

- pandas
- Hugging Face Datasets
- scikit-learn
- Hugging Face Transformers

Dataset:

- IMDB reviews: https://huggingface.co/datasets/imdb