Introduction to Modern NLP

Mayank Singh

ACM-IKDD Summer School on Data Science, IIT Gandhinagar, 7th July 2022

Mayank Singh

Assistant Professor
CSE, IITGN

Research Interests: Code-mixing, Poisoning, Scientific Data Representations

Email: singh.mayank@iitgn.ac.in
Webpage: https://mayank4490.github.io/
Research group: https://labs.iitgn.ac.in/lingo/

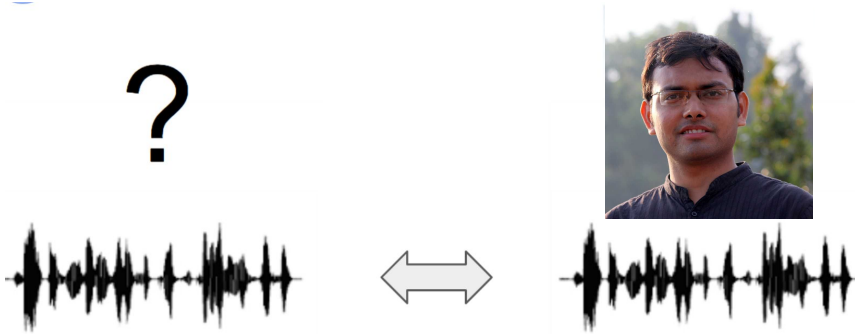
Natural Language Processing: What?

- It deals with all forms of **SPOKEN** and **WRITTEN DATA**

Natural Language Processing: What?

- It deals with all forms of **SPOKEN** and **WRITTEN** **DATA**

ASR systems

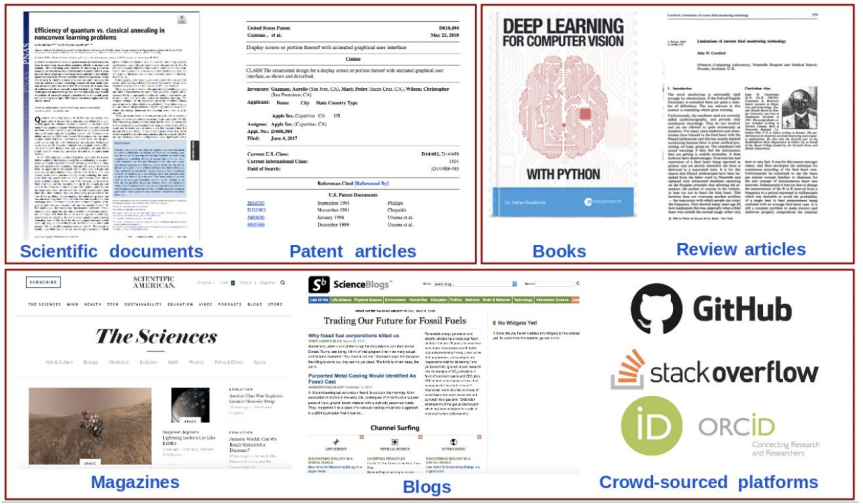


Speaker Identification Systems

Natural Language Processing: What?

- It deals with all forms of SPOKEN and **WRITTEN DATA**

Social Media



Scholarly Media

Natural Language Processing: What?

- It deals with all forms of **SPOKEN** and **WRITTEN DATA**
- A combination of both. **Any example?**

Natural Language Processing: What?

- It deals with all forms of **SPOKEN** and **WRITTEN DATA**
- A combination of both. **Any example?**

Keeping things simple: We shall focus on **Written Data** only

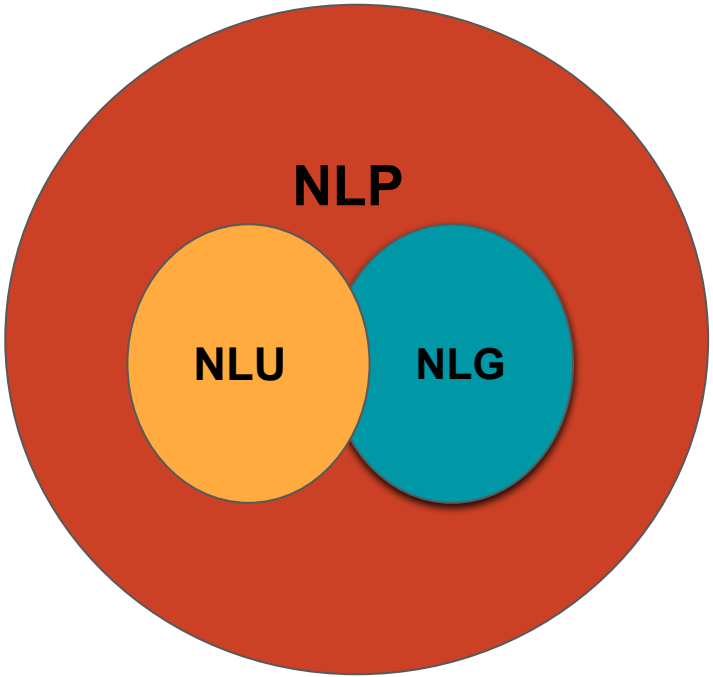
Natural Language Processing: What?

- It deals with all forms of **SPOKEN** and **WRITTEN DATA**
- A combination of both. **Any example?**

Enable computers to understand human language

The Two Components of NLP

Natural Language Understanding



Natural Language Generation

Source: https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/

NLU: Natural Language Understanding

Comprehension by computers of the **structure and meaning** of human language (e.g., English, Spanish, Japanese), allowing users to interact with the computer using natural sentences

Source: https://www.gartner.com/en/information-technology/glossary/nlu-natural-language-understanding

NLU: Natural Language Understanding

Comprehension by computers of the **structure and meaning** of human language (e.g., English, Spanish, Japanese), allowing users to interact with the computer using natural sentences

Example Block

1. Alice is swimming against the **current**.
2. The **current** version of the report is in the folder.

Source: https://www.gartner.com/en/information-technology/glossary/nlu-natural-language-understanding

NLU: Natural Language Understanding

Comprehension by computers of the **structure and meaning** of human language (e.g., English, Spanish, Japanese), allowing users to interact with the computer using natural sentences

Example Block

1.	Alice is swimming against the **current**.	**NOUN**
2.	The **current** version of the report is in the folder.	**ADJECTIVE**

Source: https://www.gartner.com/en/information-technology/glossary/nlu-natural-language-understanding

NLG: Natural Language Generation

While NLU focuses on computer reading comprehension, NLG enables computers to **write like Humans**

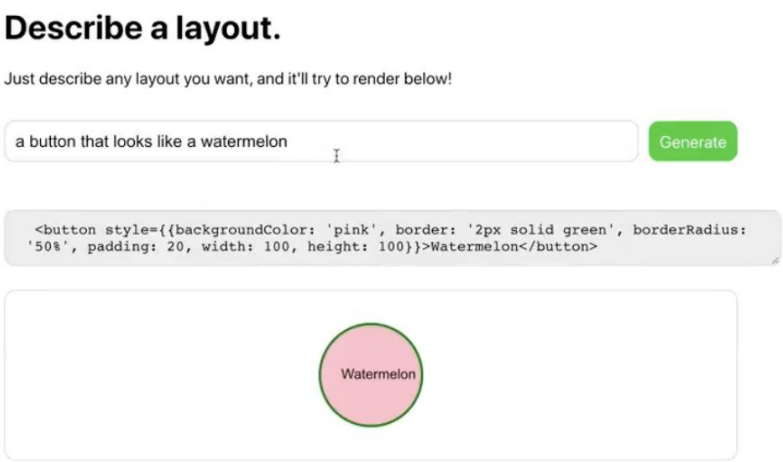
NLG: Natural Language Generation

While NLU focuses on computer reading comprehension, NLG enables computers to **write like Humans**

Example Block

Input: Layout description

Output: A JSX code



Source: https://twitter.com/sharifshameem/status/1282676454690451457

NLU Example Tasks: Text Classification

Example Block

Input:

Worst battery

Reviewed in India on 7 November 2020

Worst battery performance.
Iphone 11 is far better den this..
In 4 hour battery will come down from 100 to 15 percent.
Please dont buy this product at this price.

Expected label: Negative

Sentiment Analysis

Source: https://www.amazon.in/product-reviews/B08L5TNJHG/ref=acr_dp_hist_1?ie=UTF8&filterByStar=one_star&reviewerType=all_reviews#reviews-filter-bar

NLU Example Tasks: Text Classification

Example Block

Input: ★★★★☆ **Kidney as a load balancer**

Reviewed in India on 21 November 2020

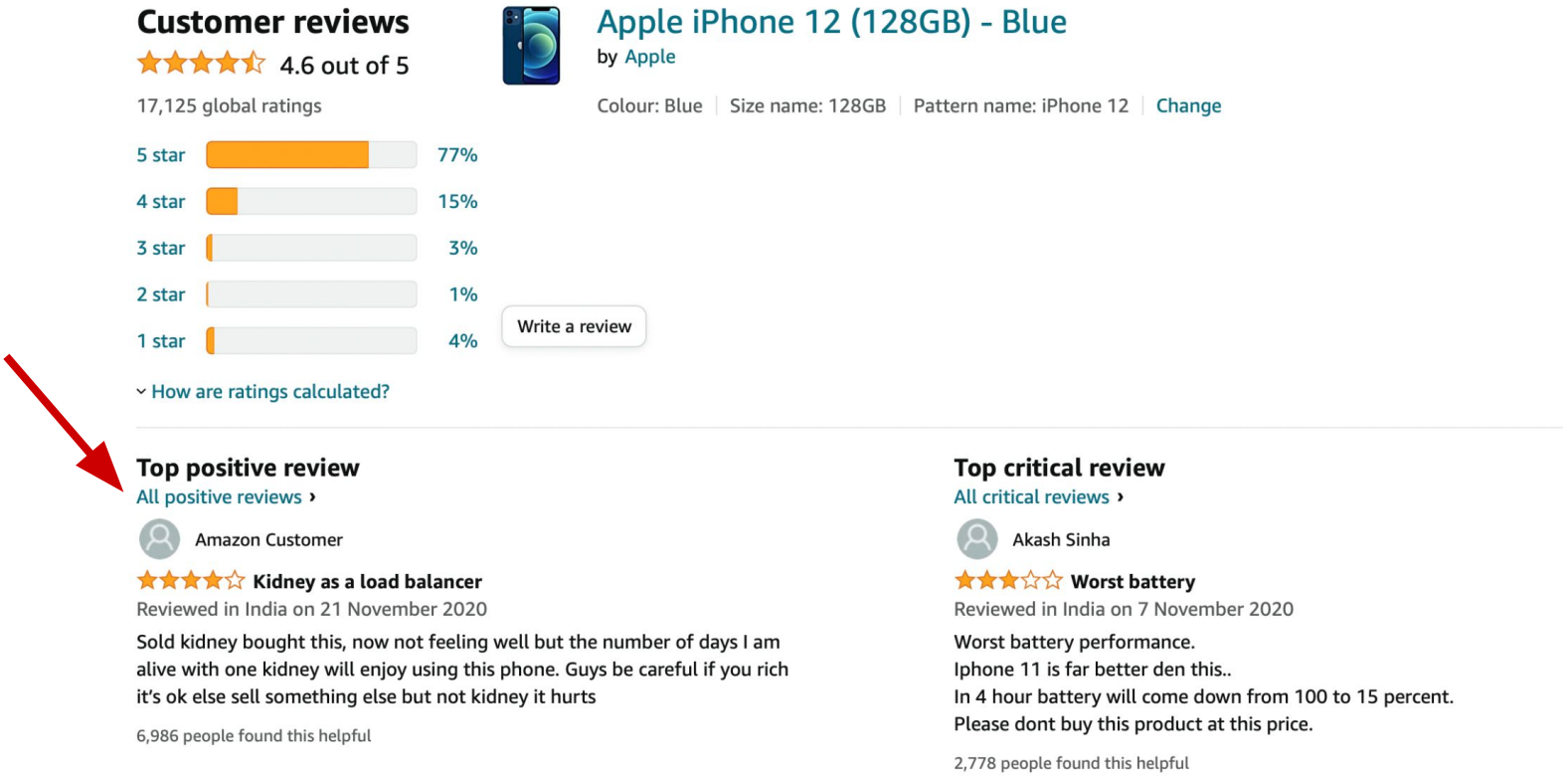
Sold kidney bought this, now not feeling well but the number of days I am alive with one kidney will enjoy using this phone. Guys be careful if you rich it's ok else sell something else but not kidney it hurts

Expected label: Negative

Sentiment Analysis

Source: https://www.amazon.in/product-reviews/B08L5TNJHG/ref=acr_dp_hist_1?ie=UTF8&filterByStar=one_star&reviewerType=all_reviews#reviews-filter-bar

NLU Example Tasks: Text Classification



Customer reviews

4.6 out of 5

17,125 global ratings

5 star 77%
4 star 15%
3 star 3%
2 star 1%
1 star 4%

How are ratings calculated?

Apple iPhone 12 (128GB) - Blue
by Apple

Colour: Blue | Size name: 128GB | Pattern name: iPhone 12 | Change

Write a review

Top positive review

All positive reviews ›

Amazon Customer

Kidney as a load balancer

Reviewed in India on 21 November 2020

Sold kidney bought this, now not feeling well but the number of days I am alive with one kidney will enjoy using this phone. Guys be careful if you rich it's ok else sell something else but not kidney it hurts

6,986 people found this helpful

Top critical review

All critical reviews ›

Akash Sinha

Worst battery

Reviewed in India on 7 November 2020

Worst battery performance.
Iphone 11 is far better den this..
In 4 hour battery will come down from 100 to 15 percent.
Please dont buy this product at this price.

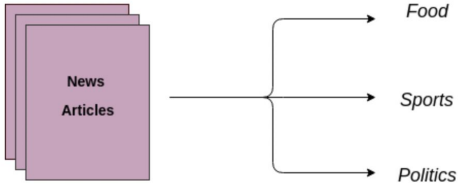
2,778 people found this helpful

Source: https://www.amazon.in/product-reviews/B08L5TNJHG/ref=acr_dp_hist_1?ie=UTF8&filterByStar=one_star&reviewerType=all_reviews#reviews-filter-bar

NLU Example Tasks: Text Classification

Spam Detection

Topic Classification



Hate Speech Detection

More Datasets: https://imerit.net/blog/17-best-text-classification-datasets-for-machine-learning-all-pbm/

NLU Example Tasks: Sentence Similarity

5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

Similarity scores with explanations and English examples from Agirre et al. (2013)

Source: https://aclanthology.org/S17-2001.pdf

NLU Benchmarks: The GLUE

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

ICLR 2019

Source: https://gluebenchmark.com/tasks

NLU Benchmarks: The GLUE

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9
2	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9
3	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6
4	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9
5	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2

Source: https://gluebenchmark.com/leaderboard

NLU Benchmarks: The SuperGLUE

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

NeurIPS 2019

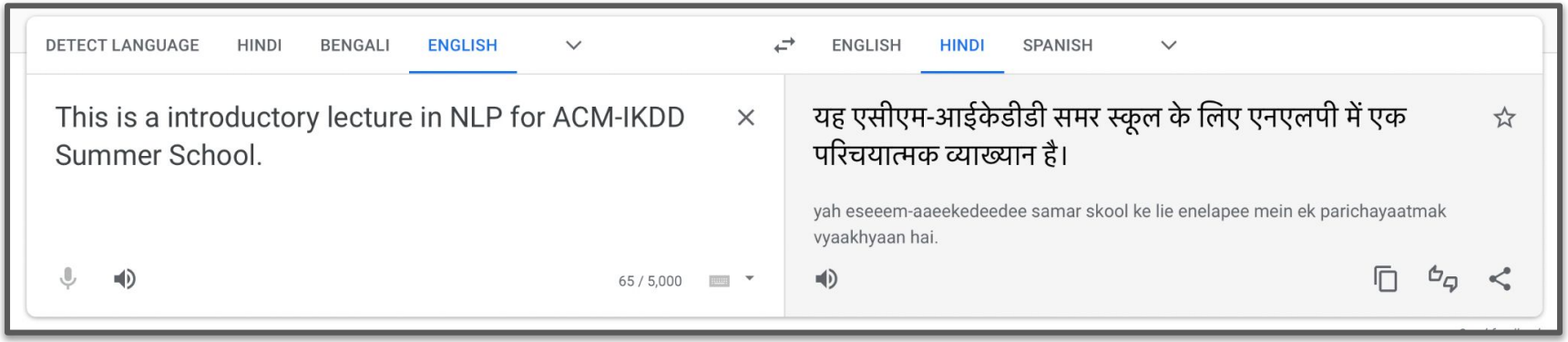
Source:https://super.gluebenchmark.com/tasks

NLU Benchmarks: The SuperGLUE

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
4	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
5	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9

Source:https://super.gluebenchmark.com/leaderboard

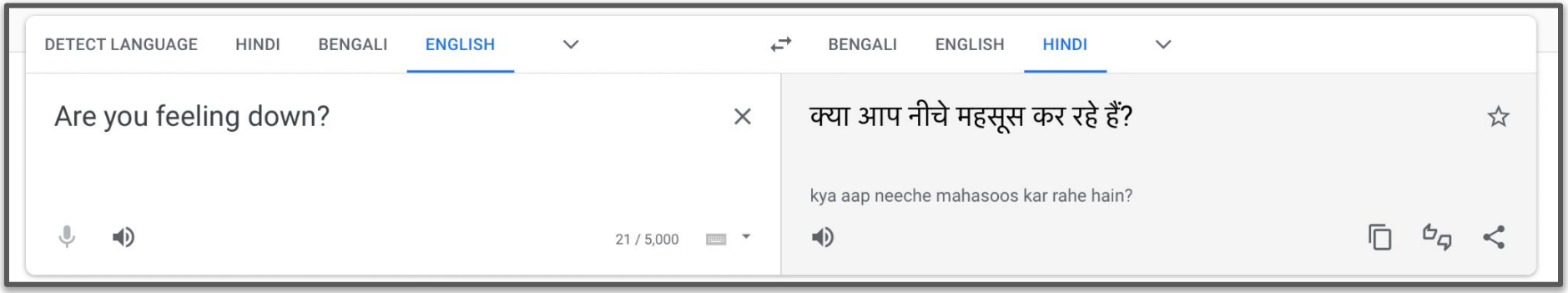
NLG Example Tasks: Machine Translation



Source Text **Target Text**

Source: https://translate.google.com/?sl=en&tl=hi&text=This%20is%20a%20introductory%20lecture%20in%20NLP%20for%20ACM-IKDD%20Summer%20School.&op=translate

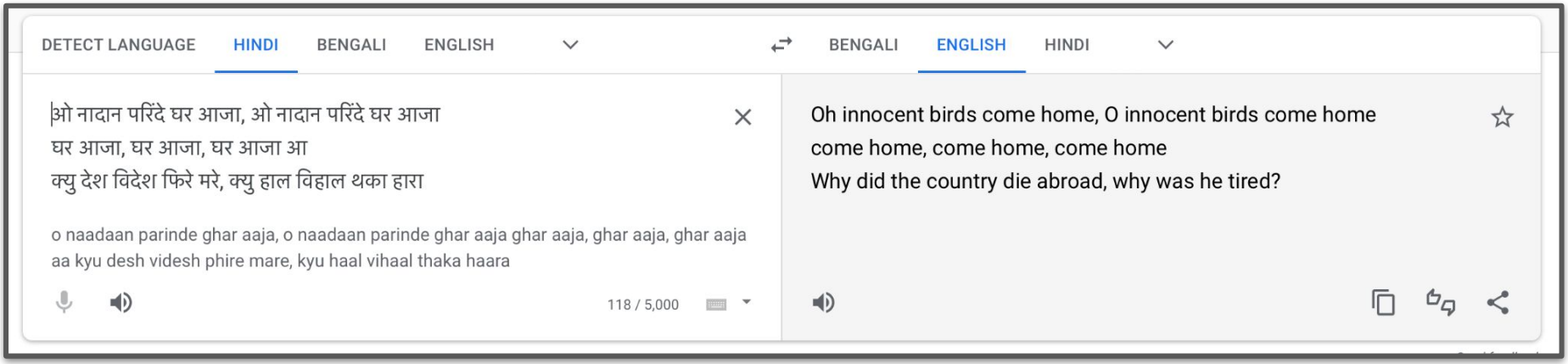
NLG Example Tasks: Machine Translation



Source Text **Target Text**

Source: https://translate.google.com/?sl=en&tl=hi&text=Are%20you%20feeling%20down%3F&op=translate

NLG Example Tasks: Machine Translation



Source Text **Target Text**

NLG Example Tasks: Summarization

Source Text	Summary Text
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

NLG Benchmark: The GEM

Dataset	Communicative Goal	Language(s)	Size	Input Type
CommonGEN (Lin et al., 2020)	Produce a likely sentence which mentions all of the source concepts.	en	67k	Concept Set
Czech Restaurant (Dušek and Jurčíček, 2019)	Produce a text expressing the given intent and covering the specified attributes.	cs	5k	Meaning Representation
DART (Radev et al., 2020)	Describe cells in a table, covering all information provided in triples.	en	82k	Triple Set
E2E clean (Novikova et al., 2017) (Dušek et al., 2019)	Describe a restaurant, given all and only the attributes specified on the input.	en	42k	Meaning Representation
MLSum (Scialom et al., 2020)	Summarize relevant points within a news article	*de/es	*520k	Articles
Schema-Guided Dialog (Rastogi et al., 2020)	Provide the surface realization for a virtual assistant	en	*165k	Dialog Act

11 Tasks

Source: https://gem-benchmark.com

NLG Benchmark: The GEM

Results: top 5 , Measures: lexical

		bleu	meteor	nist	rouge1	rouge2	rougeL	sari
data2text	common gen test							
	dart test							
	e2e nlg test	**mT5_large** mT5_xl mT5_base T5-small (Baseline) mT5_small	**mT5_large** mT5_xl mT5_base FB_NLG T5-small (Baseline)	**mT5_large** FB_NLG mT5_xl mT5_base T5-small (Baseline)	**FB_NLG** mT5_base mT5_large mT5_small mT5_xl	**mT5_base** mT5_large NUIG-DSI FB_NLG mT5_small	**NUIG-DSI** mT5_base mT5_small mT5_large mT5_xl	
	totto test	**T5-base (Baseline)** T5-xl (Baseline) ByT5-xl (Baseline) mT5_xl T5-large (Baseline)	**T5-xl (Baseline)** mT5_xl T5-base (Baseline) mT5_large T5-large (Baseline)	**T5-base (Baseline)** mT5_xl ByT5-xl (Baseline) T5-xl (Baseline) T5-large (Baseline)	**T5-base (Baseline)** T5-xl (Baseline) T5-large (Baseline) ByT5-xl (Baseline) mT5_xl	**T5-base (Baseline)** ByT5-xl (Baseline) T5-large (Baseline) T5-xl (Baseline) ByT5-base (Baseline)	**T5-base (Baseline)** ByT5-xl (Baseline) T5-xl (Baseline) T5-large (Baseline) ByT5-base (Baseline)	
	cs restaurants test	**TGen_lemma-tag+RNNLM** TGen+RNNLM TGen mT5_base mT5_xl	**TGen** TGen+RNNLM mT5_large mT5_base TGen_lemma-tag+RNNLM	**TGen_lemma-tag+RNNLM** TGen TGen+RNNLM mT5_base mT5_xl	**TGen_lemma-tag+RNNLM** TGen TGen+RNNLM ByT5-base (Baseline) mT5_base	**TGen_lemma-tag+RNNLM** TGen TGen+RNNLM mT5_base mT5_large	**TGen_lemma-tag+RNNLM** TGen TGen+RNNLM mT5_base ByT5-base (Baseline)	
	web nlg en test	**FB_NLG** mT5_xl T5-xl (Baseline) mT5_large mT5_base	**FB_NLG** mT5_xl T5-xl (Baseline) mT5_large mT5_base	**FB_NLG** mT5_xl T5-xl (Baseline) mT5_large mT5_base	**FB_NLG** T5-xl (Baseline) mT5_xl mT5_large ByT5-base (Baseline)	**FB_NLG** mT5_xl T5-xl (Baseline) ByT5-base (Baseline) ByT5-xl (Baseline)	**FB_NLG** mT5_xl ByT5-base (Baseline) T5-xl (Baseline) ByT5-xl (Baseline)	

Source: https://gem-benchmark.com

The Fundamental NLP tasks

The Fundamental NLP tasks

- **Tokenization:** Segmenting a sequence of characters into tokens

Example Block

Input: I am talking about tokenization today
Expected Output: I, am, talking, about, tokenization, today

The Fundamental NLP tasks

- **Tokenization:** Segmenting a sequence of characters into tokens

Example Block

Input: I am talking about tokenization today
Expected Output: I, am, talking, about, tokenization, today

- **Stop-word removal:** Removing common words

Example Block

Input: I, am, talking, about, tokenization, today
Expected Output: talking, tokenization, today

The Fundamental NLP tasks

- **Lemmatization:** Reducing tokens to the base forms

Example Block

Input: talking, tokenization, today
Expected Output: talk, tokenize, today

The Fundamental NLP tasks

- **Lemmatization:** Reducing tokens to the base forms

Example Block

Input: talking, tokenization, today
Expected Output: talk, tokenize, today

- **Part of Speech (PoS) tagging:** Assigning each token a particular part of speech tag, based on both its definition and its context.

Example Block

Input: talking, tokenization, today
Expected Output: Verb, Noun, Noun

The Fundamental NLP tasks

- **Named Entity Extraction:** Identifying the named entities like Person, Country, organization, etc.

Example Block

Input: I love India and its culture

Expected Output: I love India (COUNTRY) and its culture

The Fundamental NLP tasks

- **Named Entity Extraction:** Identifying the named entities like Person, Country, organization, etc.

Example Block

Input: I love India and its culture

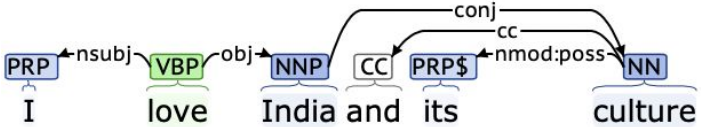
Expected Output: I love 

and its culture

- **Parse Tree Generation:**

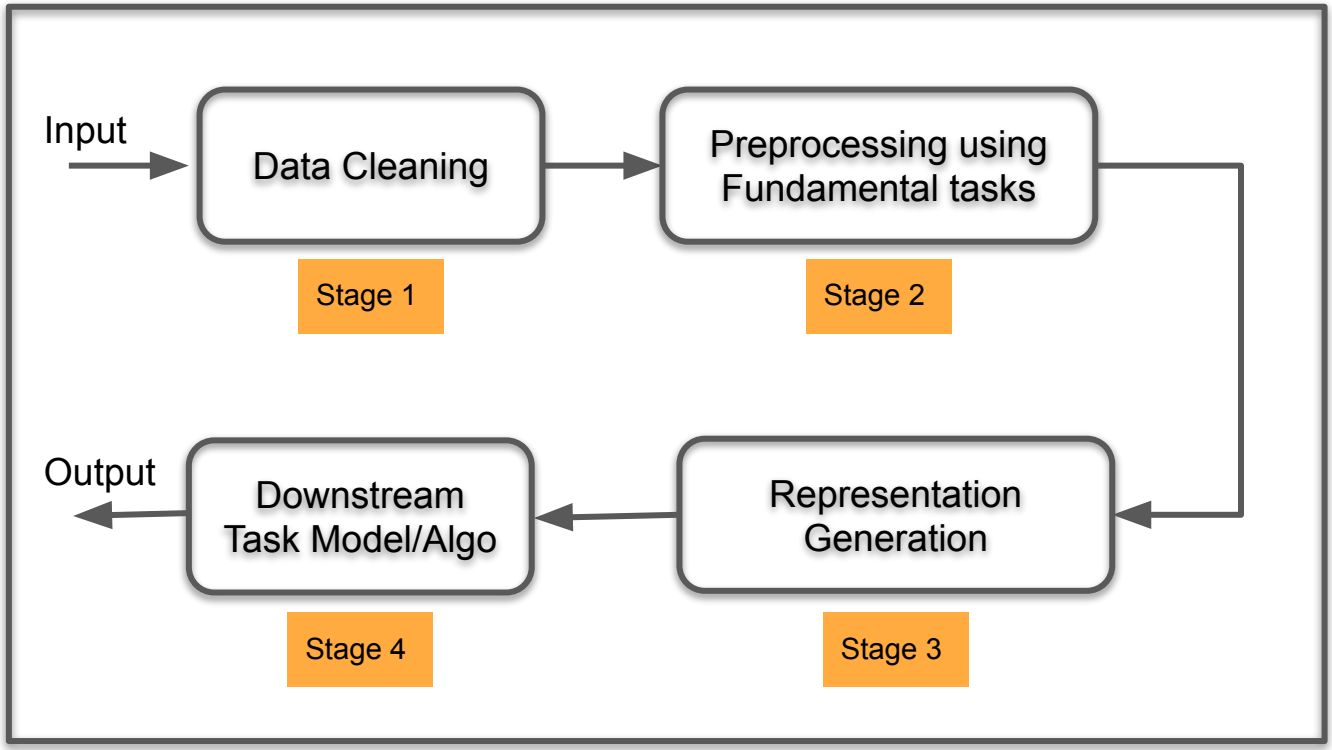
Example Block

Input: I love India and its culture

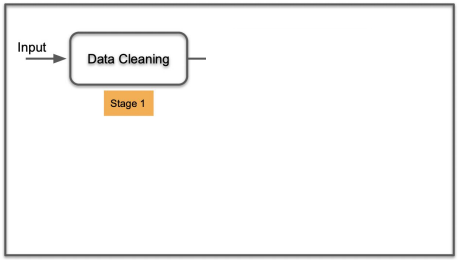
Expected Output: 

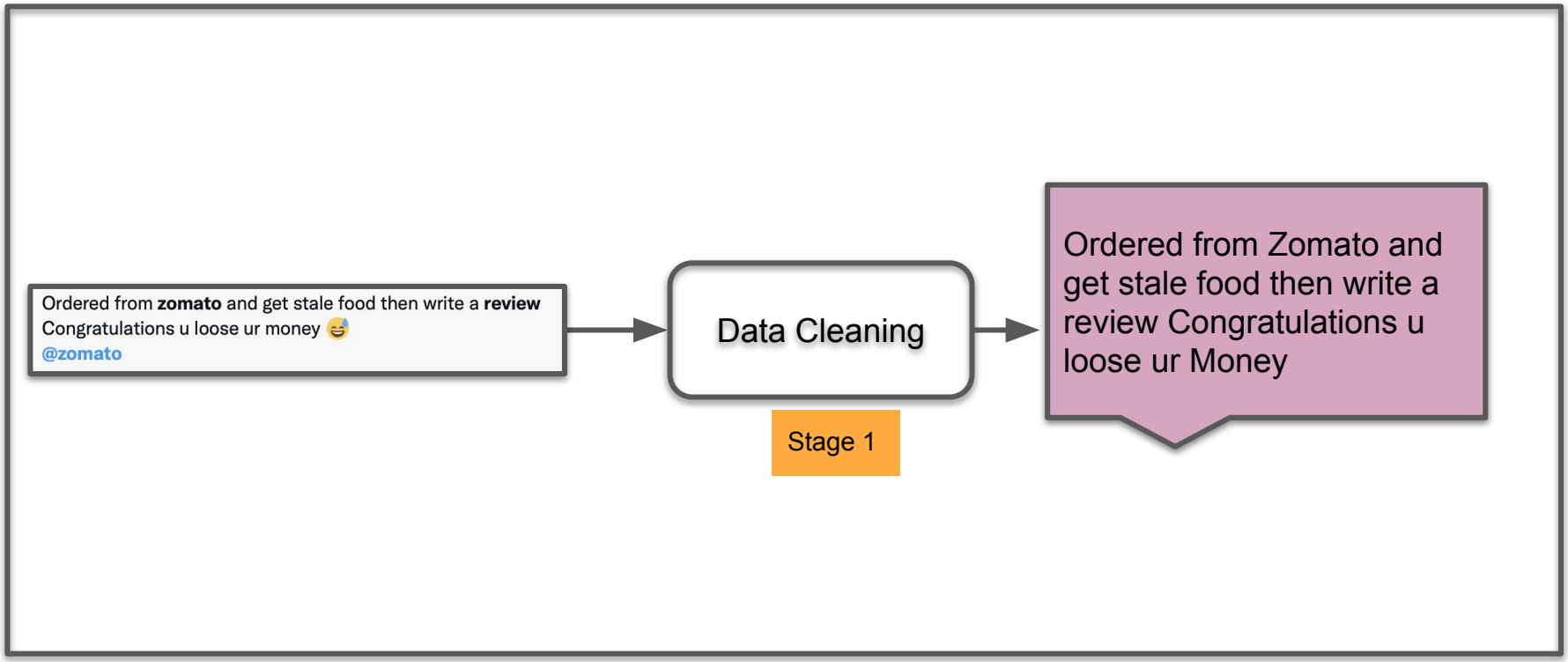
The Traditional NLP Pipeline

The Traditional NLP Pipeline

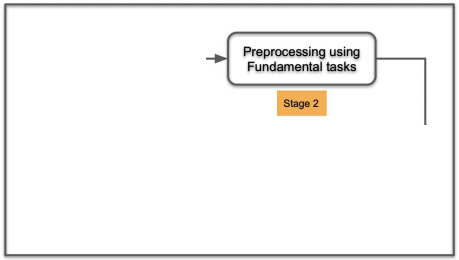


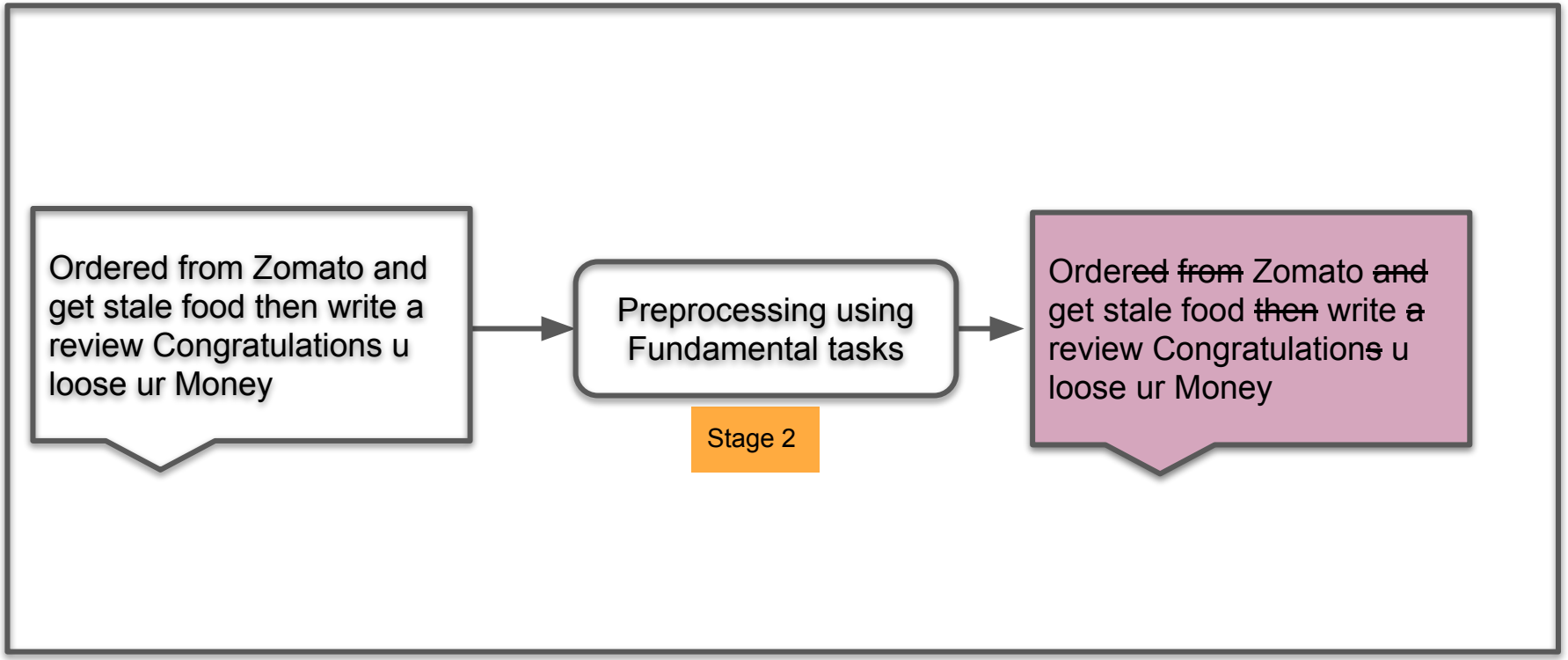
A running example: Stage 1



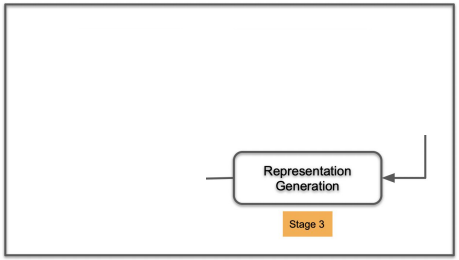


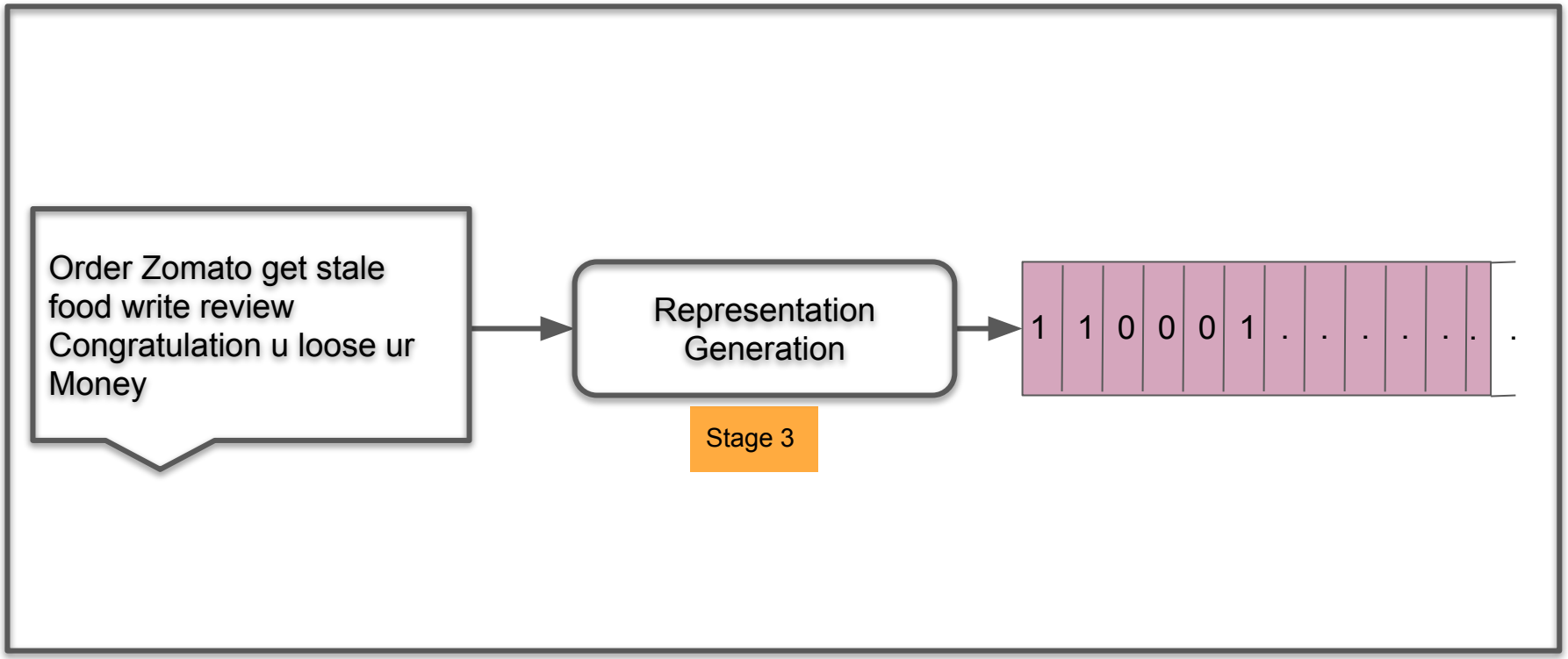
A running example: Stage 2



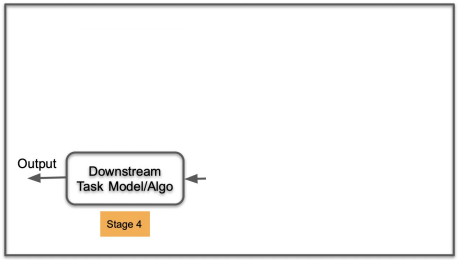


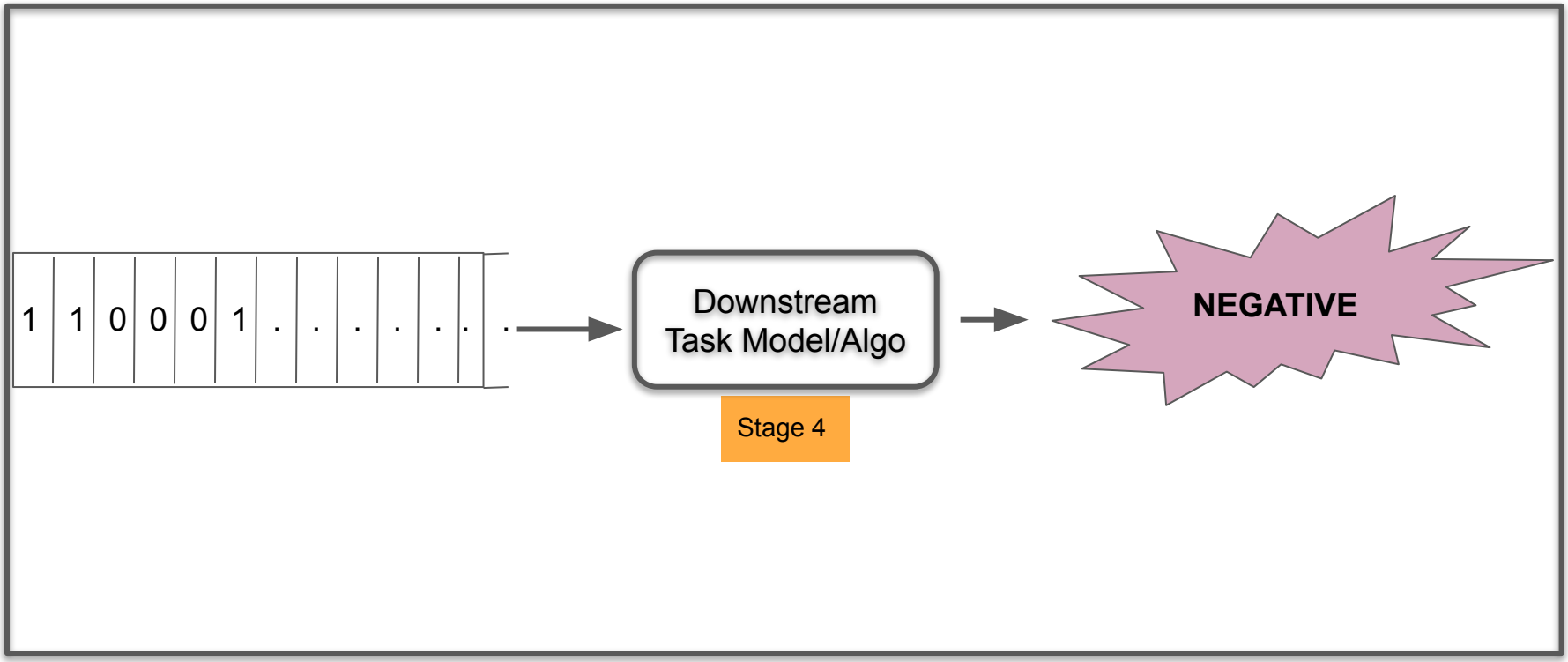
A running example: Stage 3



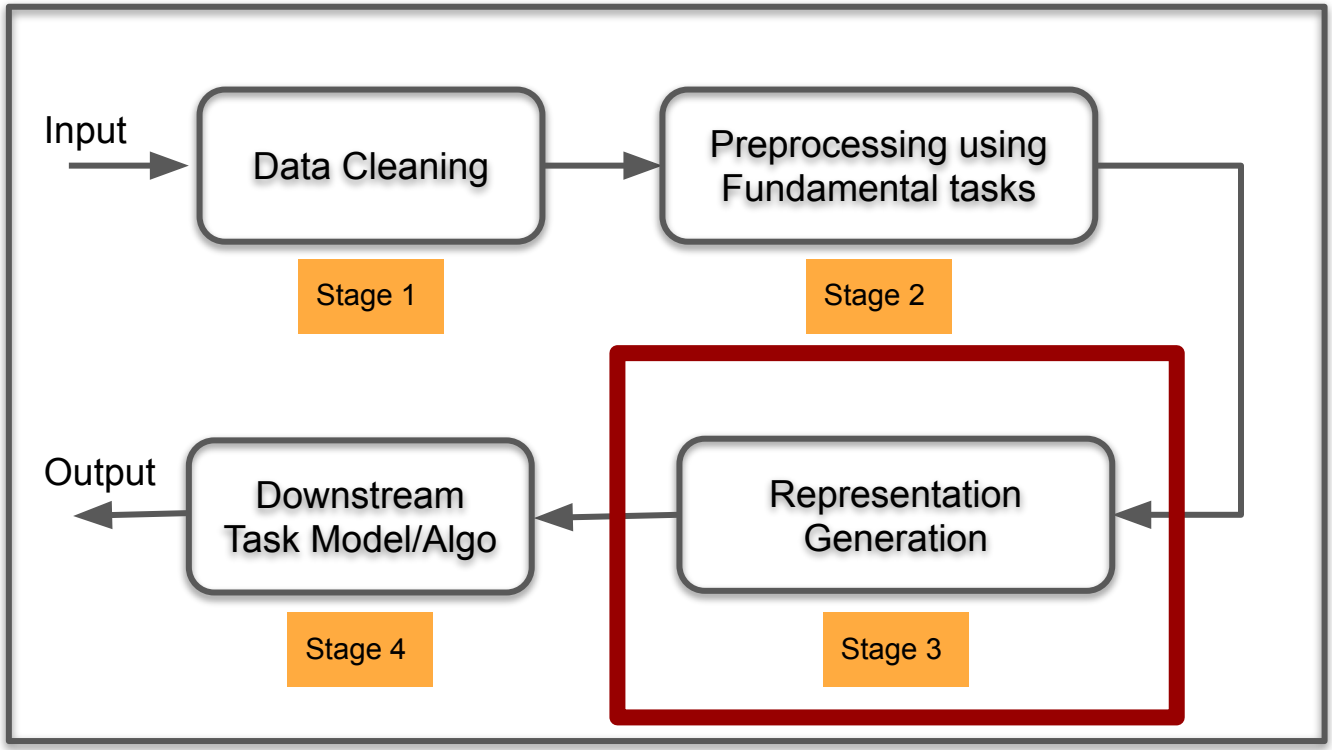


A running example: Stage 4

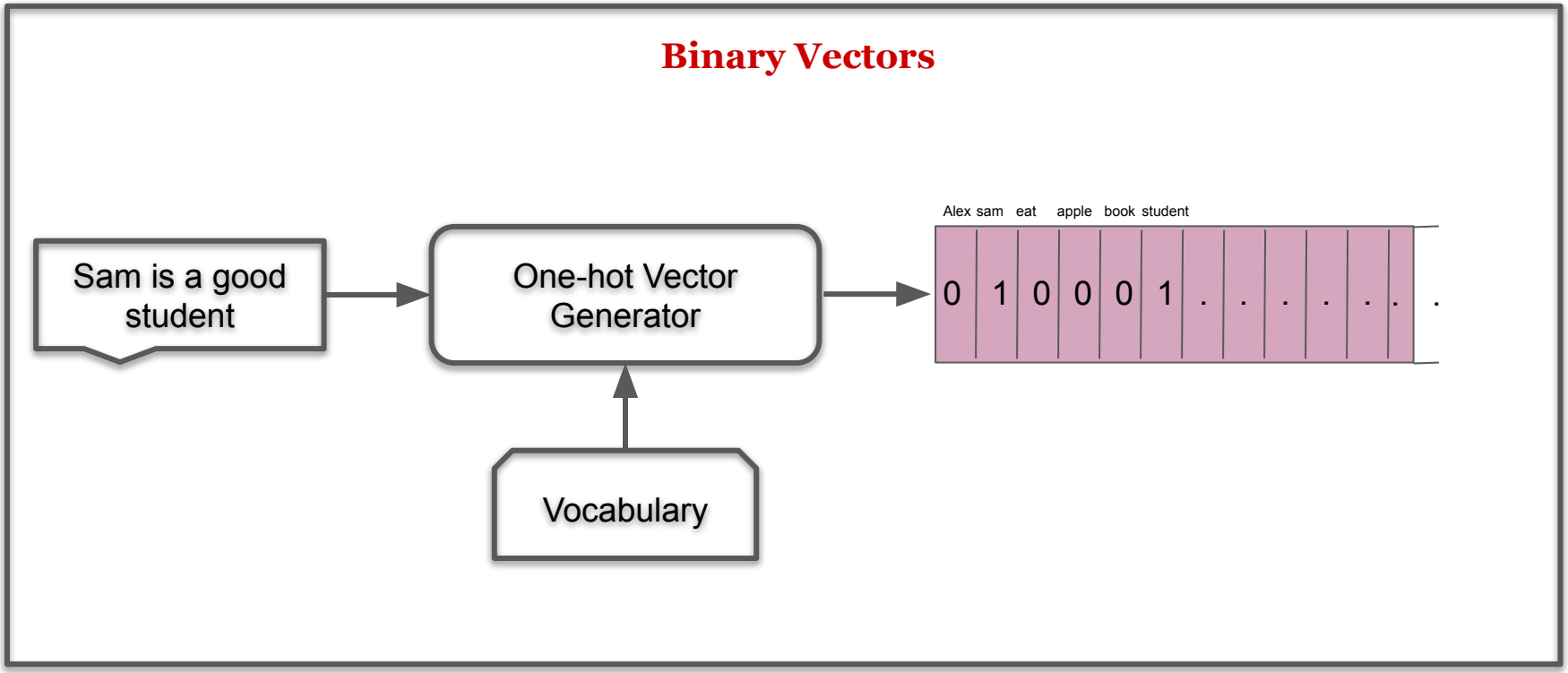




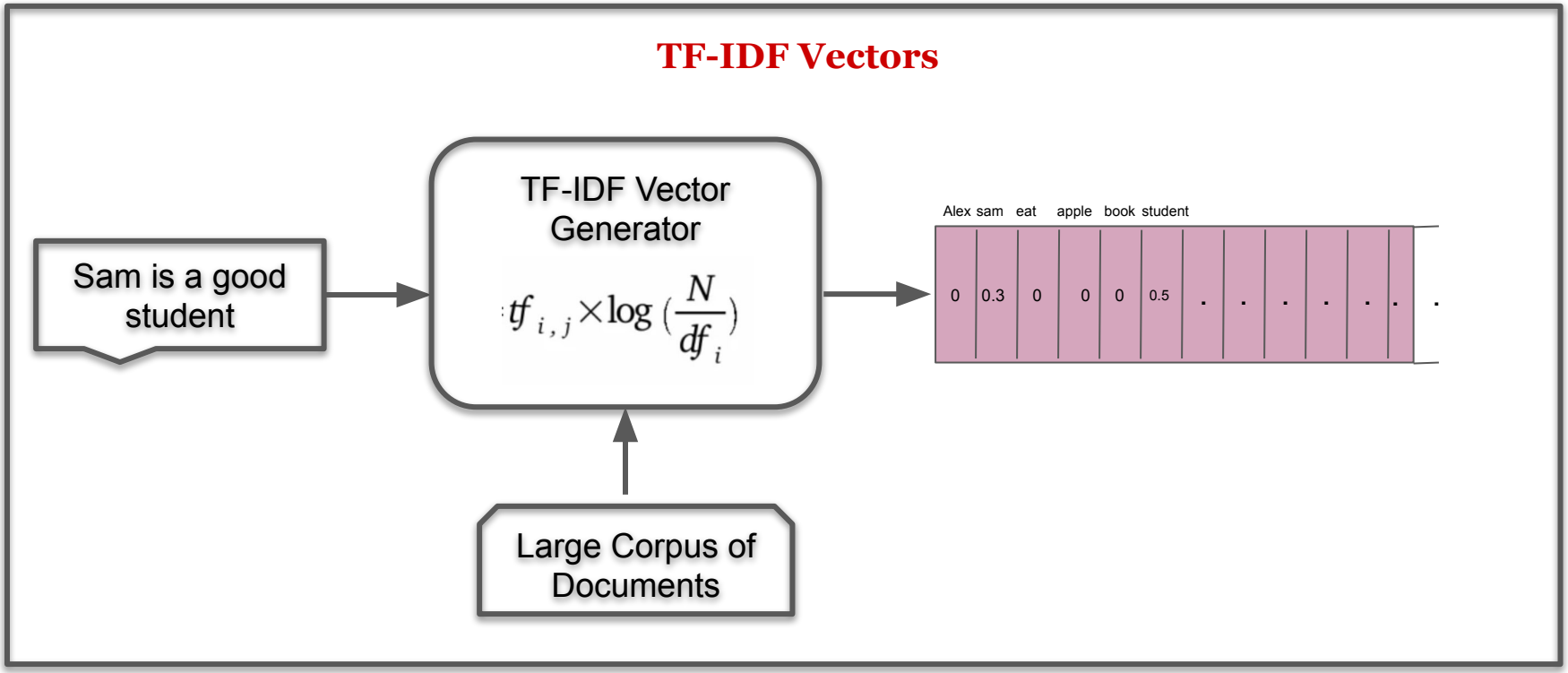
The Traditional NLP Pipeline: The Major Bottleneck



Representation Generation via Distributional Semantics



Representation Generation via Distributional Semantics



Representation Generation via Distributional Semantics

- **High Dimensional Vector**
  Size = no. of tokens in the vocabulary

- **Very Sparse**
  Non-zero dimensions <= the no. of tokens in the sentence

- **Based on bag-of-words (BoW) assumption, which does not captures:**
  - **position in text**
  - **semantics**
  - **co-occurrences in different documents**

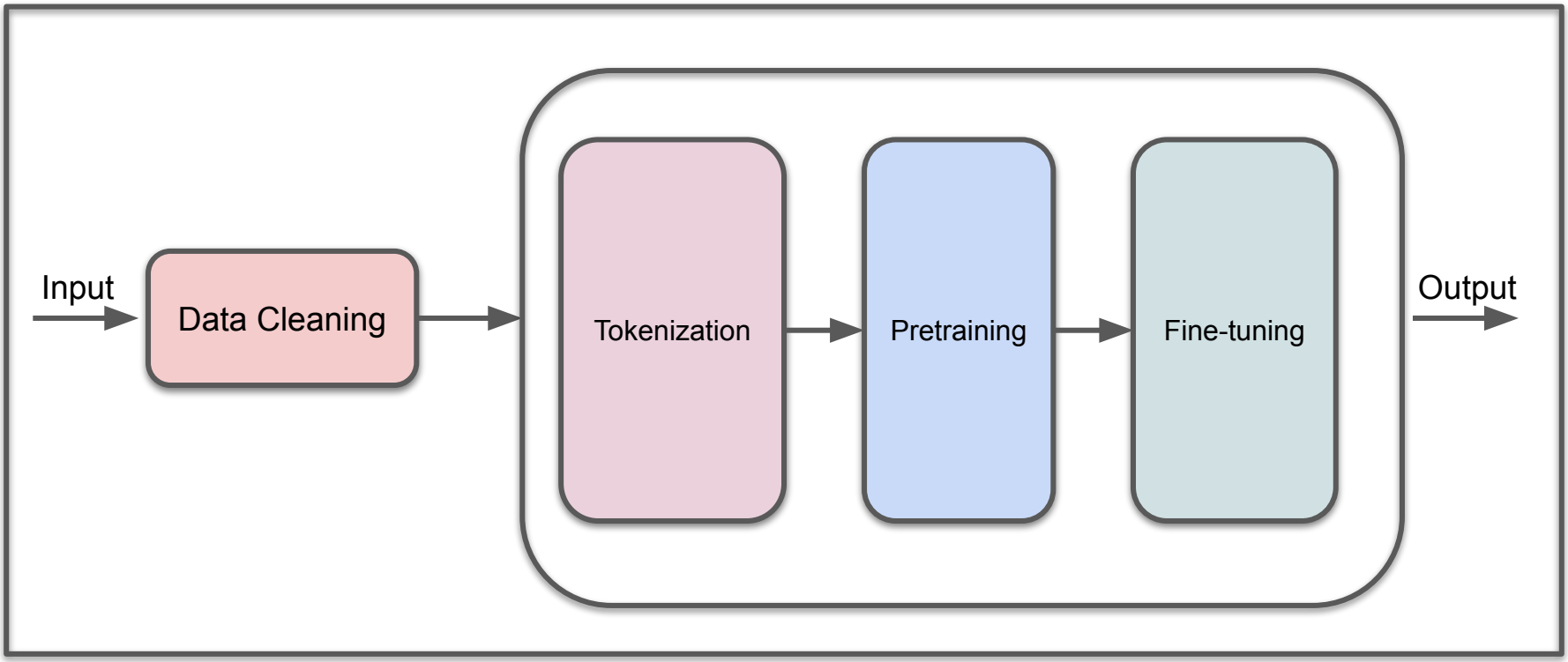
Emergence of Deep Learning for NLP

- **How to create concise length vectors?**
  Length of vectors < 1000

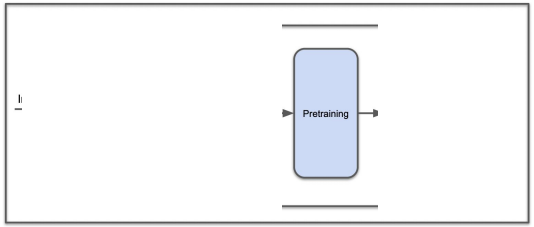
- **End-to-end training**
  No need to explicitly extract syntactic and semantic features

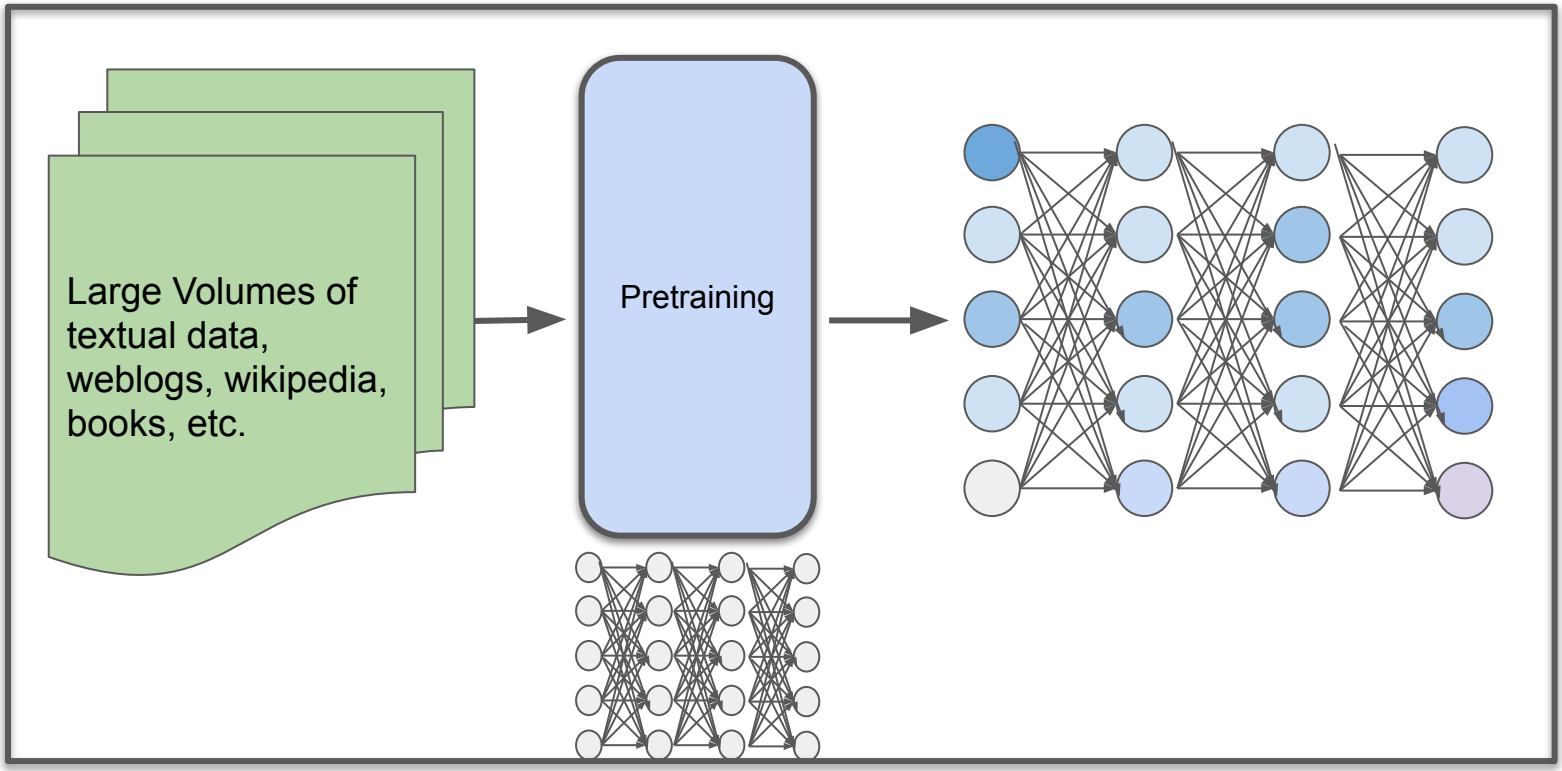
- **Dividing the training into two phases:**
  - **General purpose learning**
  - **Task specific learning**

The DL-based NLP Pipeline

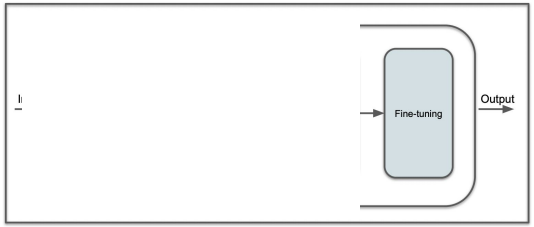


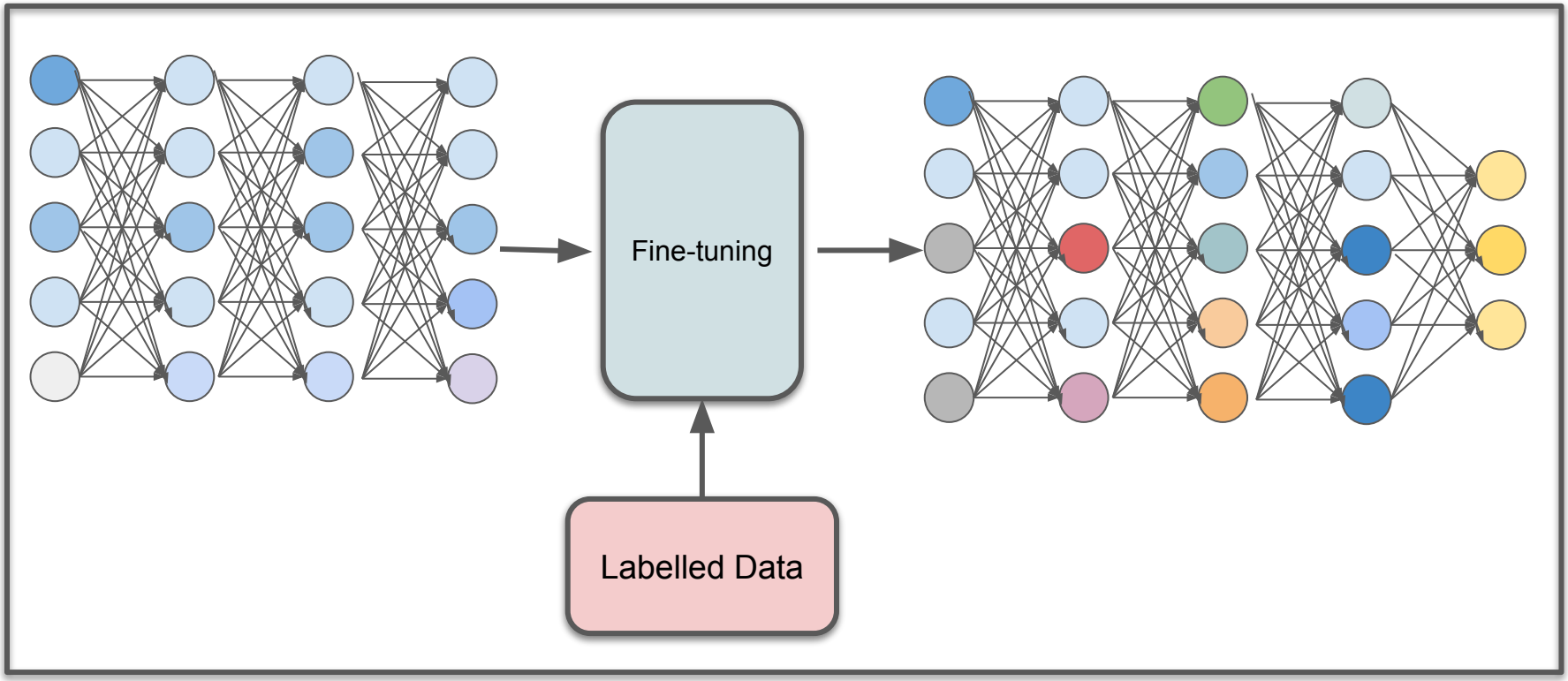
The DL-based NLP Pipeline: Pretraining





The DL-based NLP Pipeline: Fine-tuning





More details in the next lecture….

Email: singh.mayank@iitgn.ac.in
Webpage: https://mayank4490.github.io/