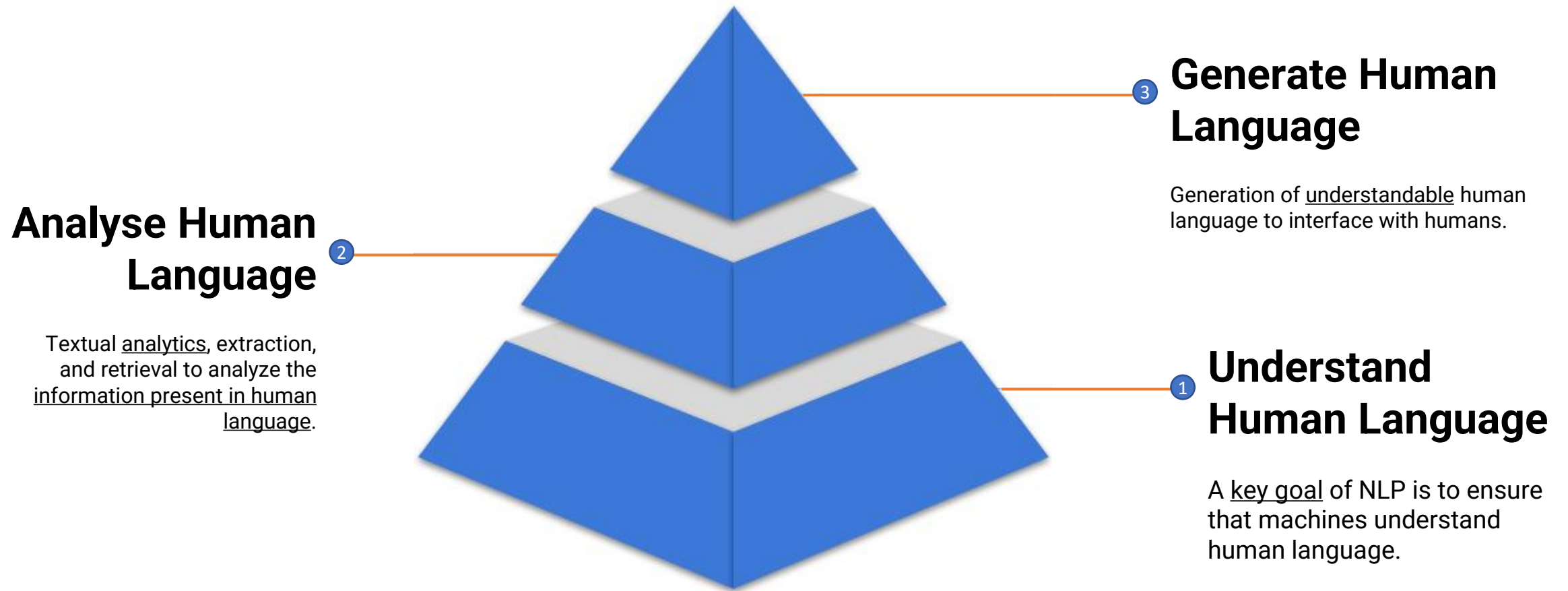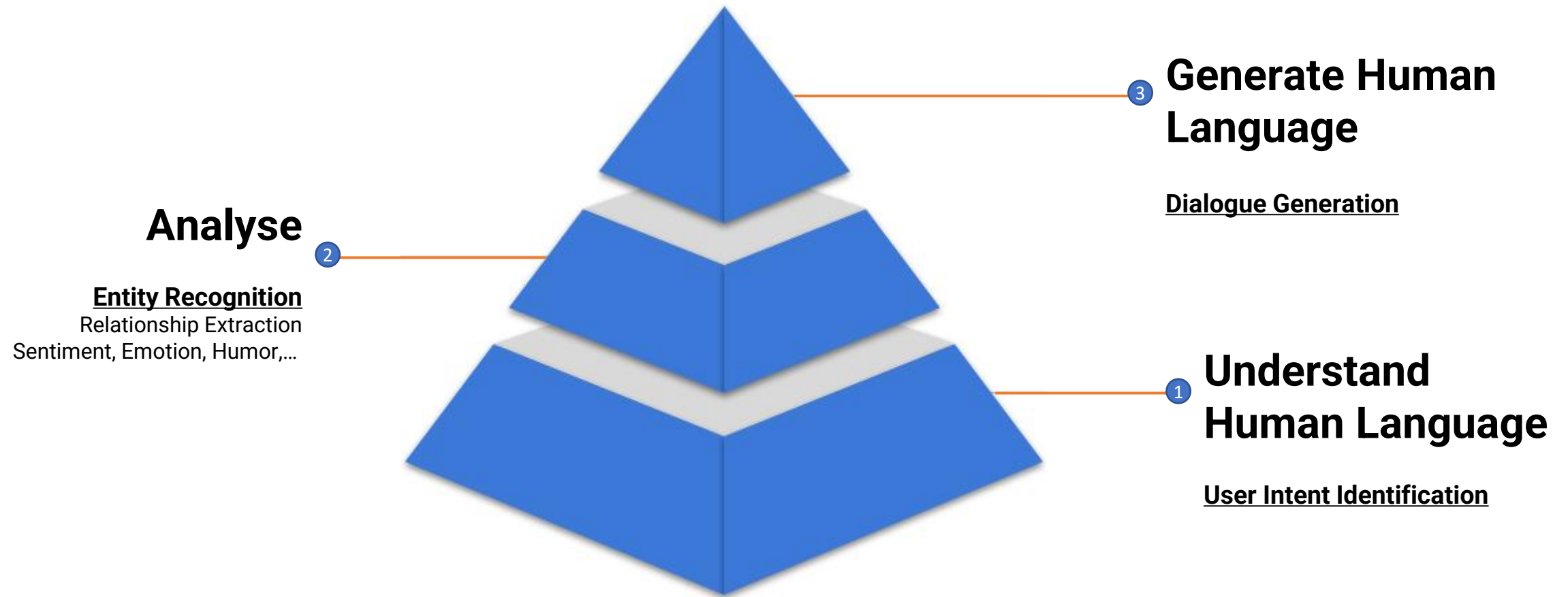# Dialogue Processing:
The Role of NLP and its Building Blocks

Dr Diptesh Kanojia
Surrey Institute for People-Centred AI
Department of Computer Science, University of Surrey

UNIVERSITY OF SURREY

People-Centred AI
UNIVERSITY OF SURREY

Natural Language Processing (NLP)

Generate Human Language

3

Generation of <u>understandable</u> human language to interface with humans.

Analyse Human Language

2

Textual <u>analytics</u>, extraction, and retrieval to analyze the <u>information present in human language</u>.

Understand Human Language

1

A <u>key goal</u> of NLP is to ensure that machines understand human language.

Natural Language Processing (NLP): A Dialogue Perspective

Generate Human Language

3

Dialogue Generation

Analyse

2

Entity Recognition
Relationship Extraction
Sentiment, Emotion, Humor,...

Understand Human Language
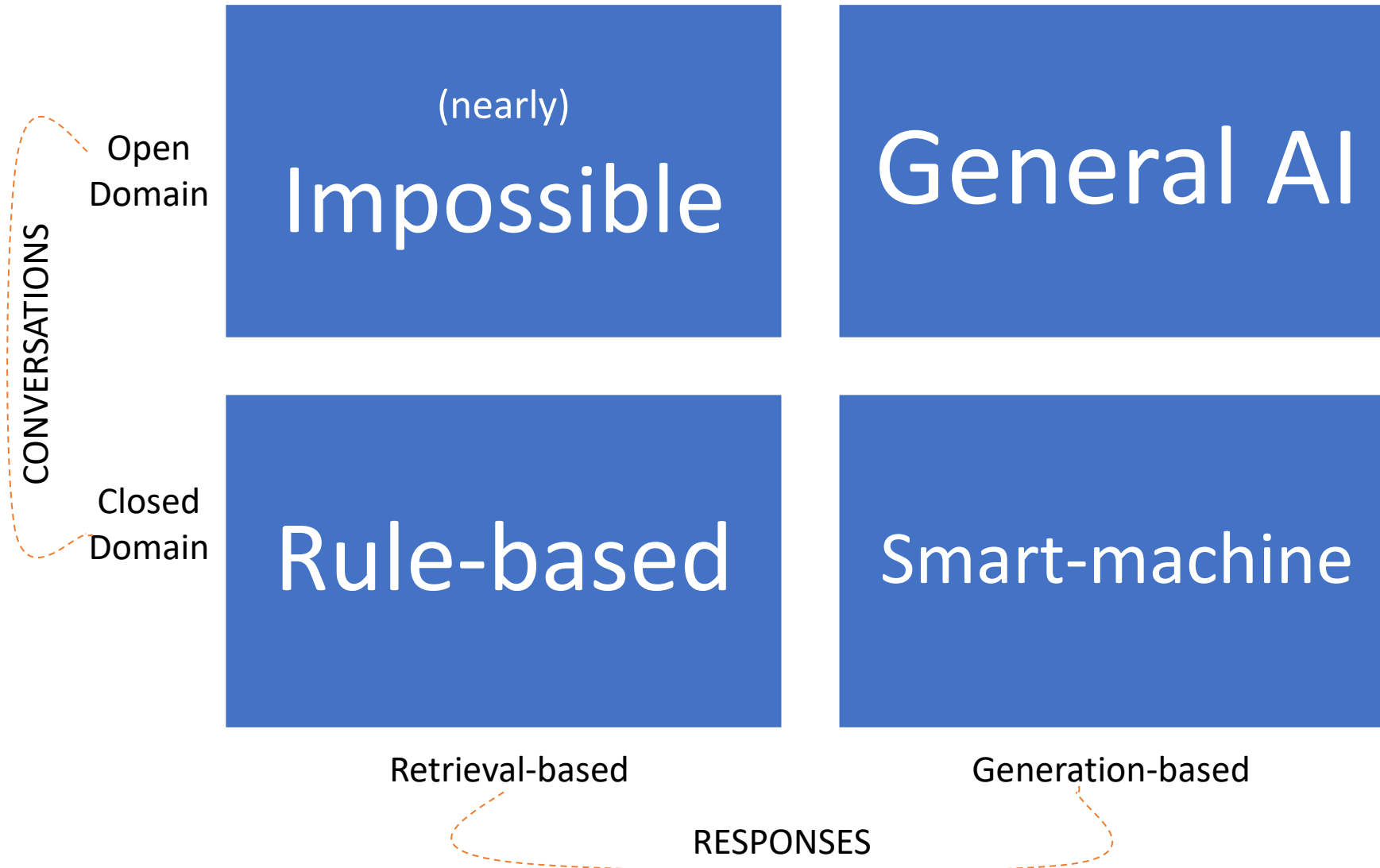
1

User Intent Identification

UNIVERSITY OF SURREY

Why Conversational AI?

- Text Messaging is a popular form of interaction and chatbots are able to streamline interaction between people and 'services'.

- Chatbots are scalable

- Always available – less dependence of human resources.
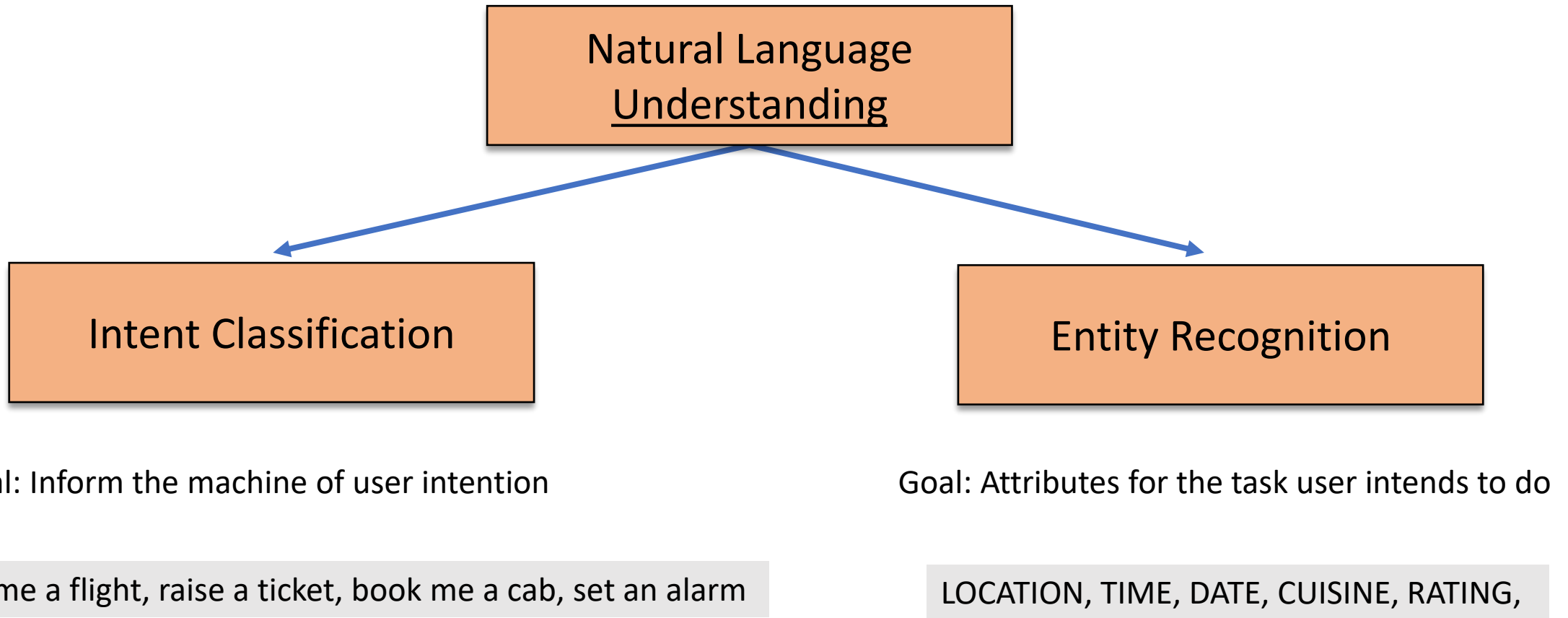
- Helpful for organization in multiple geographies.

Conversational Framework
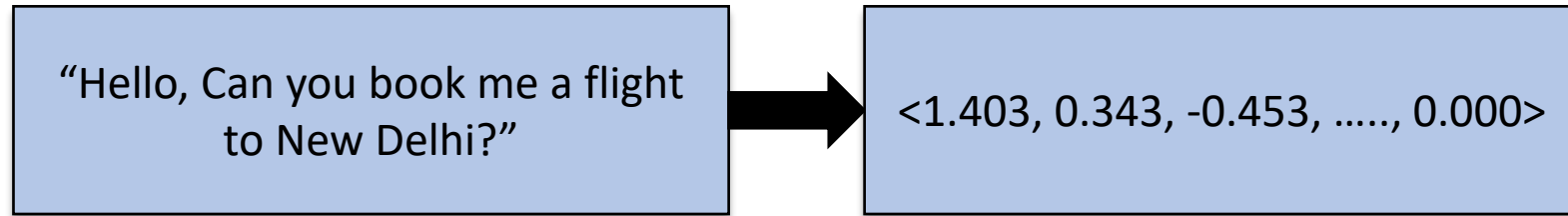
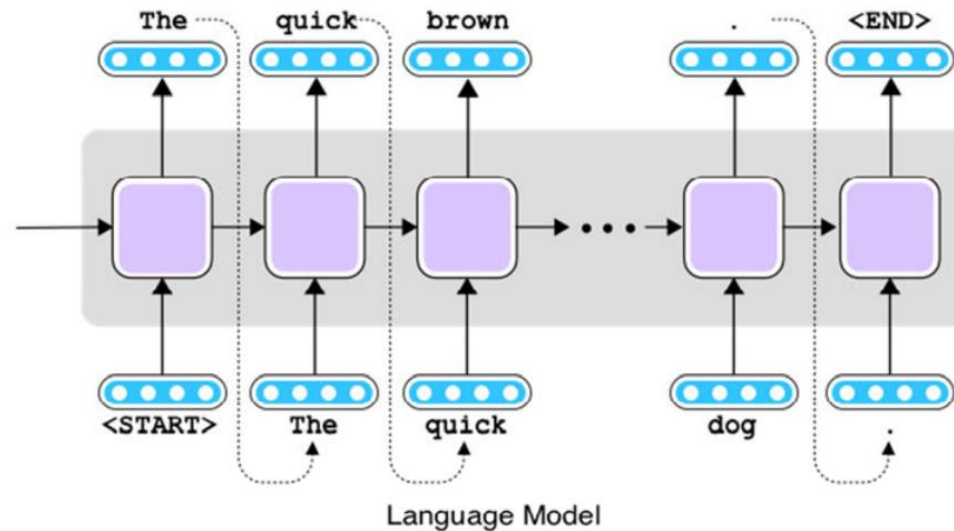Dialogue System Architecture: Research Domains
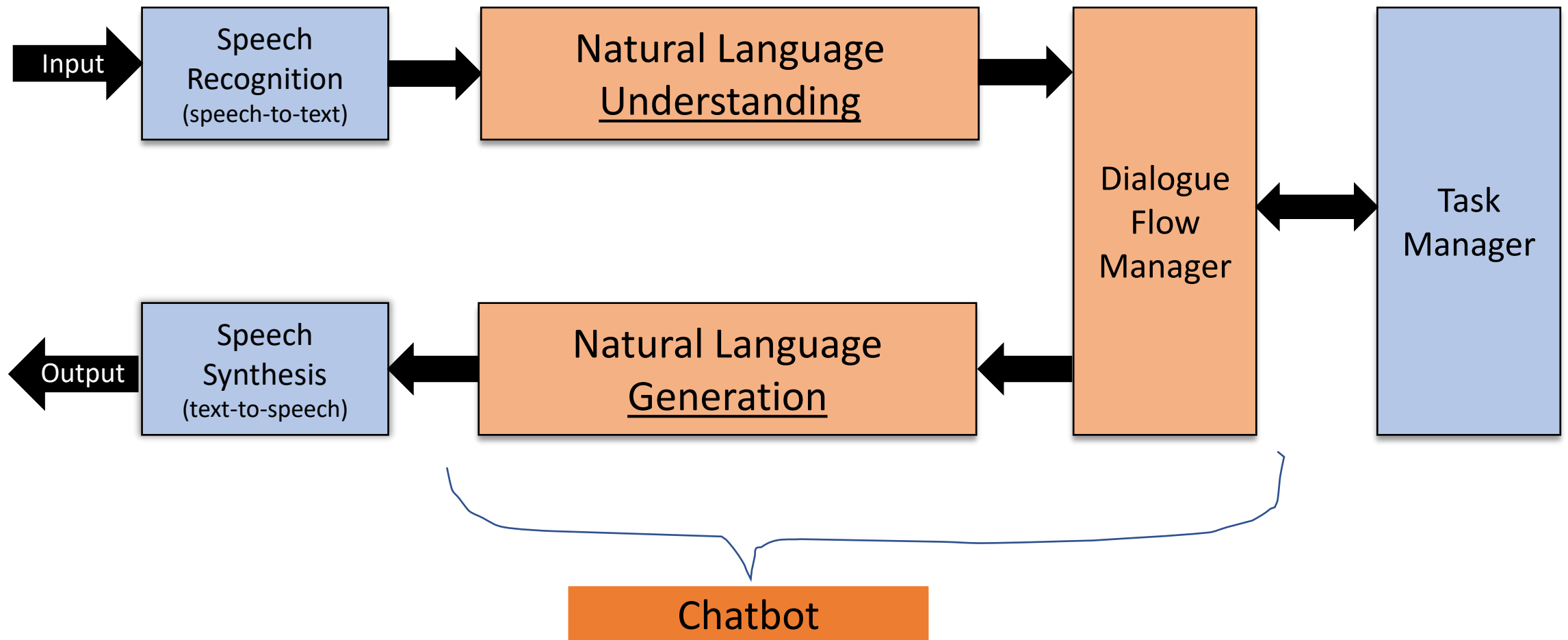
NLU Engine Components
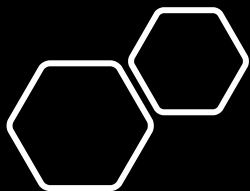
Language Understanding

- Machine Understanding

"Hello, Can you book me a flight to New Delhi?" ➔ <1.403, 0.343, -0.453, ..., 0.000>

- Language Models
  - Tokenization
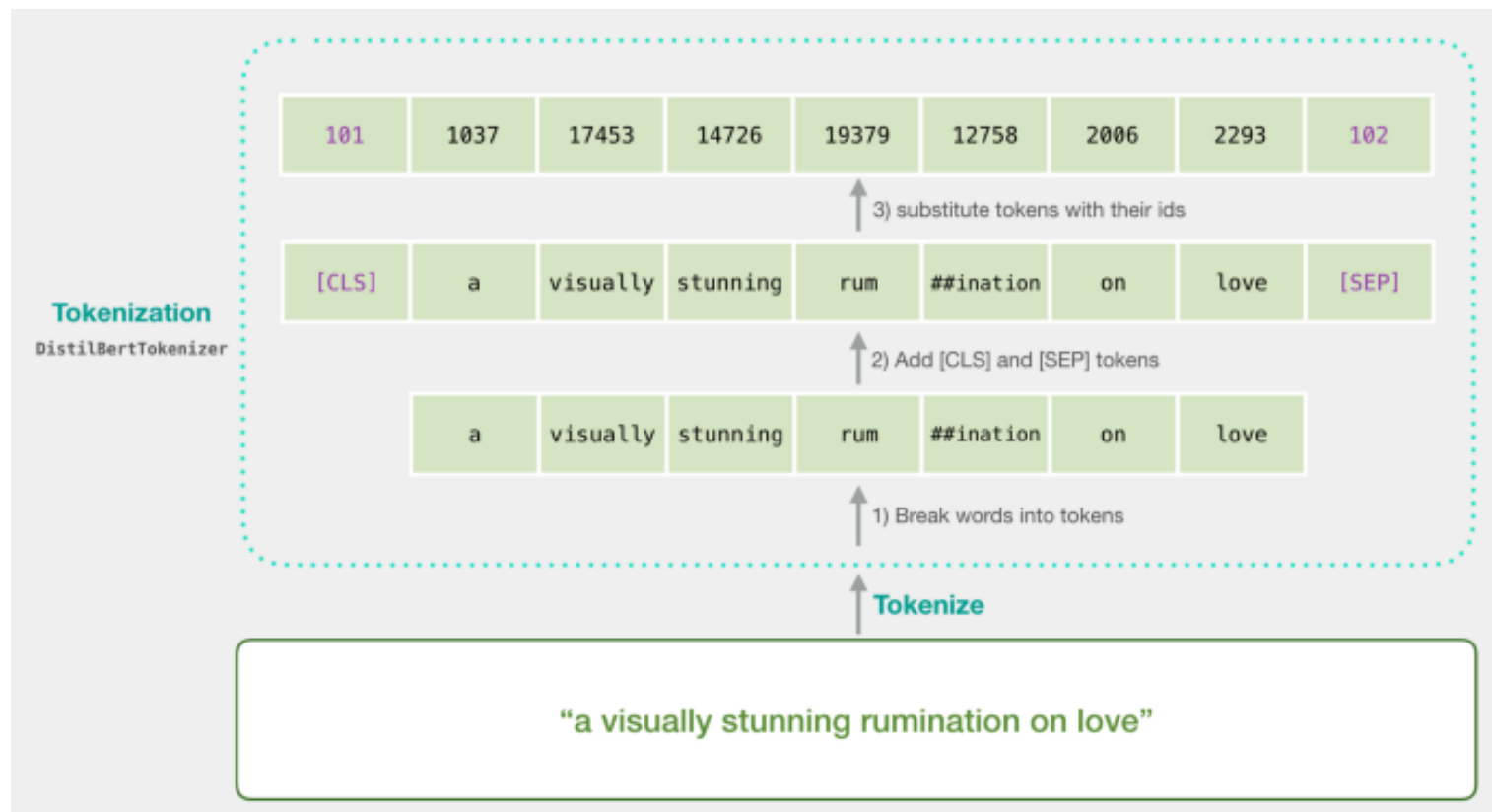 - Vectorization

Language Model

Tokenization

- Allows the model to understand <u>known tokens</u>
- <u>Breaks unknown tokens into sub-words</u>
- <u>Sub-words based on partial known words</u>

101	1037	17453	14726	19379	12758	2006	2293	102

↑ 3) substitute tokens with their ids

Tokenization
DistilBertTokenizer

[CLS]	a	visually	stunning	rum	##ination	on	love	[SEP]

↑ 2) Add [CLS] and [SEP] tokens

	a	visually	stunning	rum	##ination	on	love	

↑ 1) Break words into tokens

↑ **Tokenize**

"a visually stunning rumination on love"

Word Embeddings: Vectorization

- Word embeddings are a powerful concept that can give your application a boost in the performance, but they cannot take context into account (words can have many meanings)

- Contextualised embeddings or simply contextualisation, is a way to transform the entire sentence into a series of vectors that take into the context. Notable attempts in contextualised embeddings include CoVe3 and ELMo, although the biggest breakthrough was achieved by BERT, a Transformer-based pretrained language model.

- BERT contextualises the input through a series of Transformer encoder layers, so it inherits all the strengths of the Transformer. Its self-attention mechanism enables it to "random access" over the input and capture long-term dependencies among input tokens. So unlike traditional language models, the Transformer can take into account the context in both directions


Semi-supervised Learning Step

Model:

Dataset:

Objective: Predict the masked word (langauge modeling)

Attention!

- It is like having some sort of key-value store that contains all of the input's information and looking it up with a query (in the context).

- The stored values are not just a single vector, but a list of vectors for each token associated with its corresponding keys.
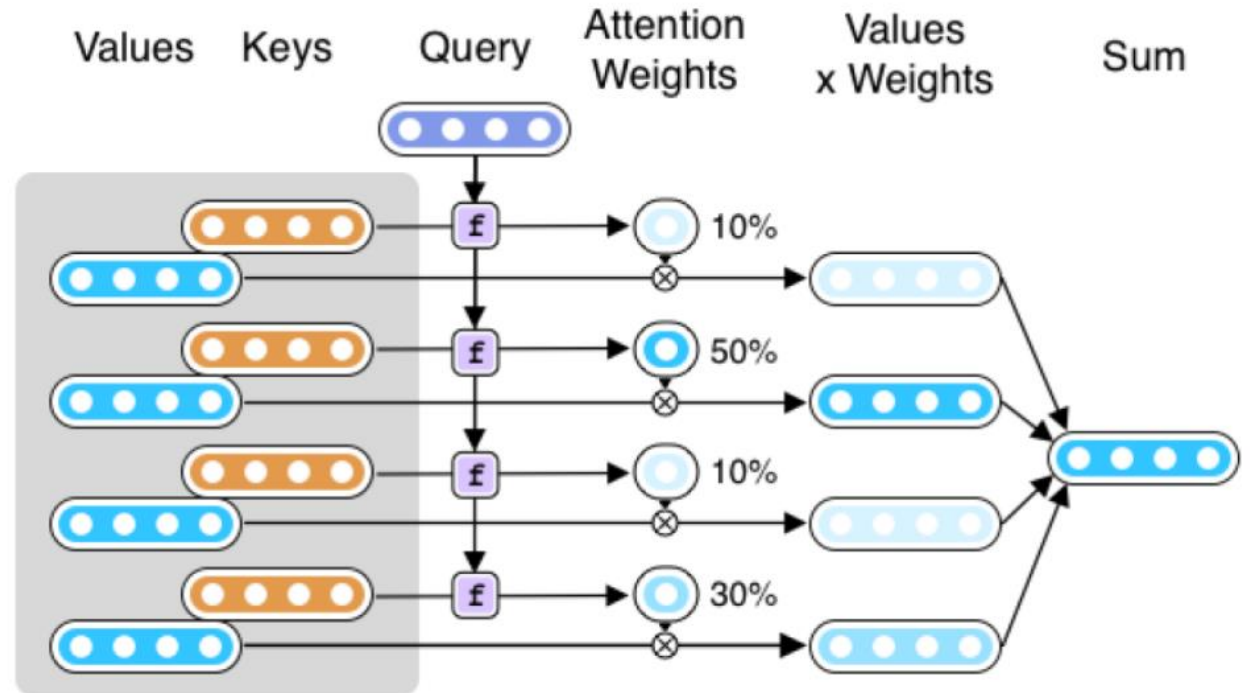
Intent Classification: Task Perspective

- In essence, an intent classifier analyses texts automatically and categorizes them into intents such as booking, spam, complaint, query_reservation, query_service, query_complaint, and so on.

- It enables the organization to be more customer-centric, particularly in areas like customer service and sales.

- Leads to faster dealing with massive number of queries while providing individualized service.
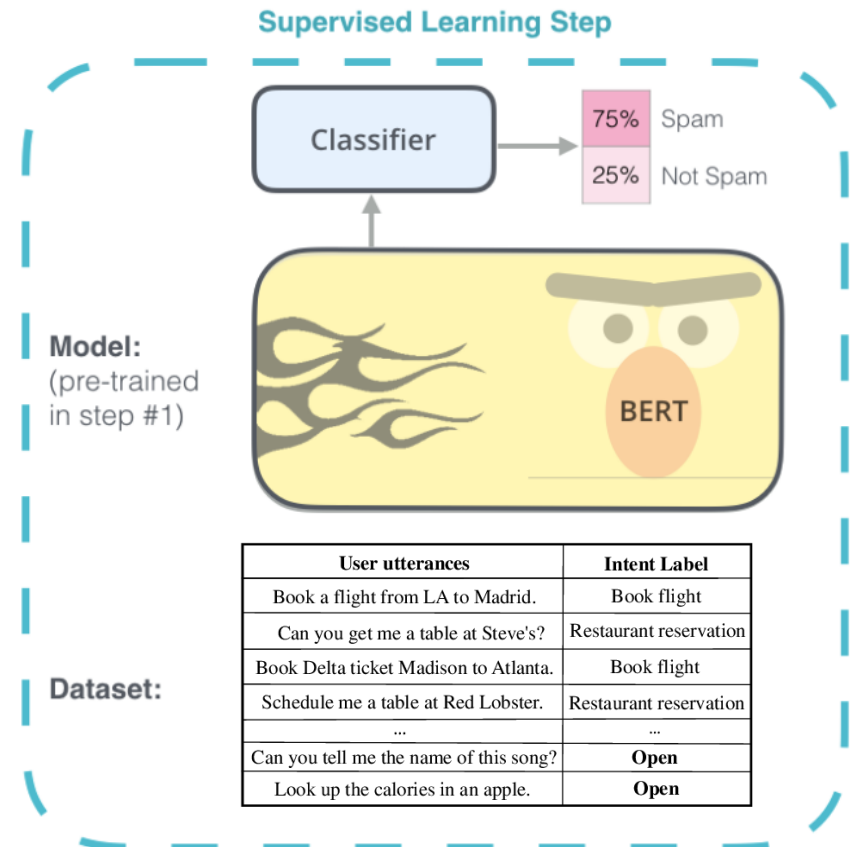
Intent Classification: Dataset

• Potential text coupled with intent as a label.

User utterances	Intent Label
Book a flight from LA to Madrid.	Book flight
Can you get me a table at Steve's?	Restaurant reservation
Book Delta ticket Madison to Atlanta.	Book flight
Schedule me a table at Red Lobster.	Restaurant reservation
...	...
Can you tell me the name of this song?	**Open**
Look up the calories in an apple.	**Open**

Classification Task

- Each data instance is converted into a vector

- The machine learns to map the label to this vector

- The complete training data is mapped to its lables – instance by instance (or batch by batch)

- Machine understands which tokens correspond to what label

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Classifier → 75% Spam / 25% Not Spam

Model: (pre-trained in step #1)

BERT

Dataset:

User utterances	Intent Label
Book a flight from LA to Madrid.	Book flight
Can you get me a table at Steve's?	Restaurant reservation
Book Delta ticket Madison to Atlanta.	Book flight
Schedule me a table at Red Lobster.	Restaurant reservation
...	...
Can you tell me the name of this song?	**Open**
Look up the calories in an apple.	**Open**

Testing / Evaluation

- Held-out set from the data known as Test set.

- Evaluation using statistical measure known as F1-score which is based on Precision and Recall scores.

- Precision – quality of the prediction of intent
  - What proportion of intent identifications was actually correct?
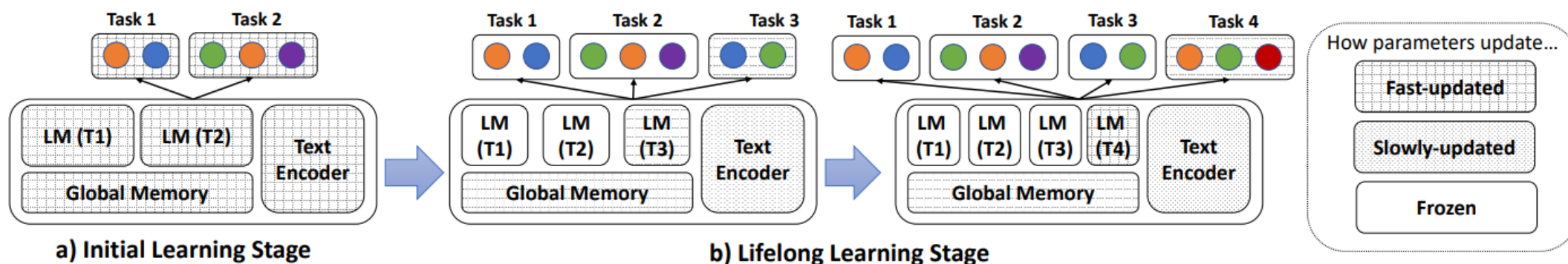
- Recall – how many of the intents were found?
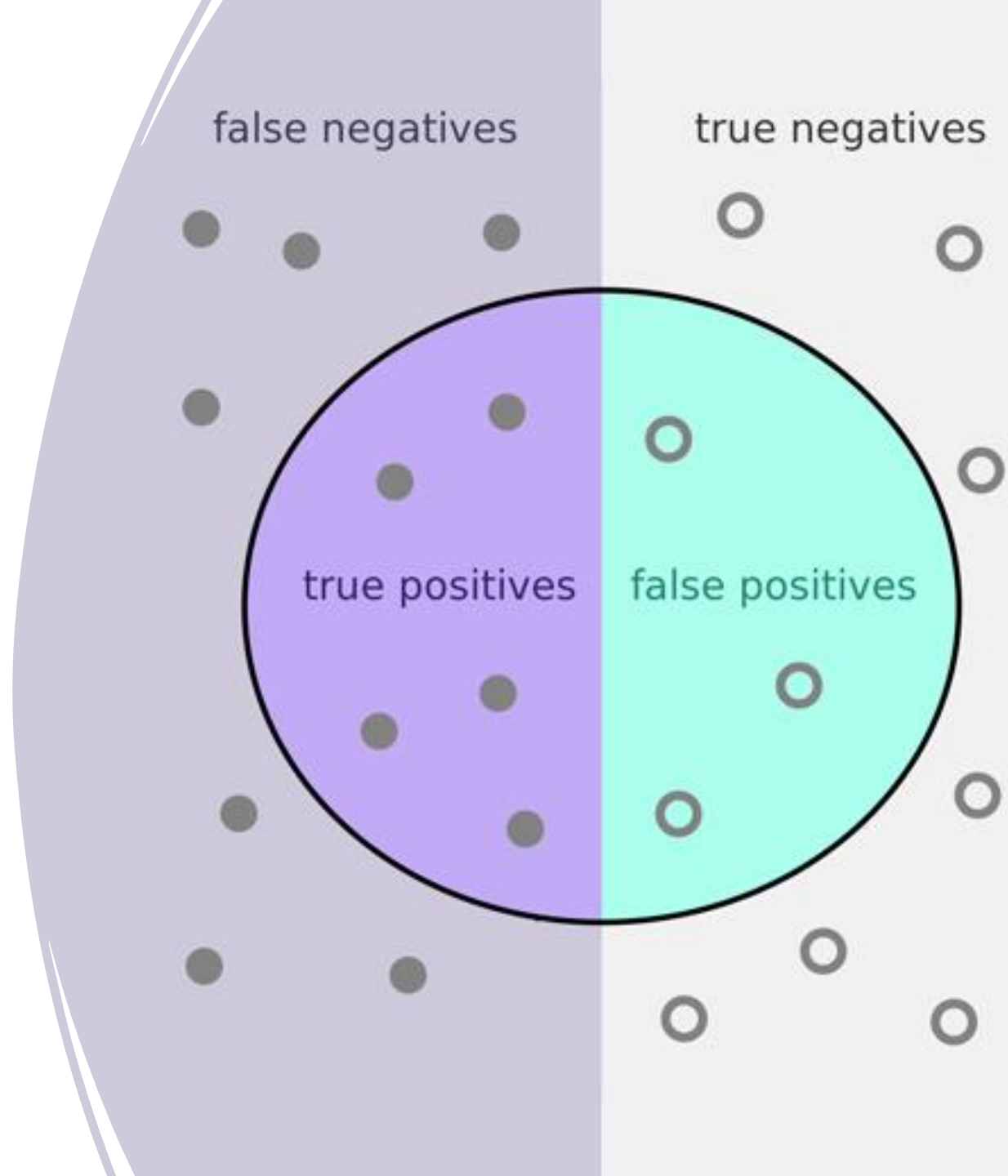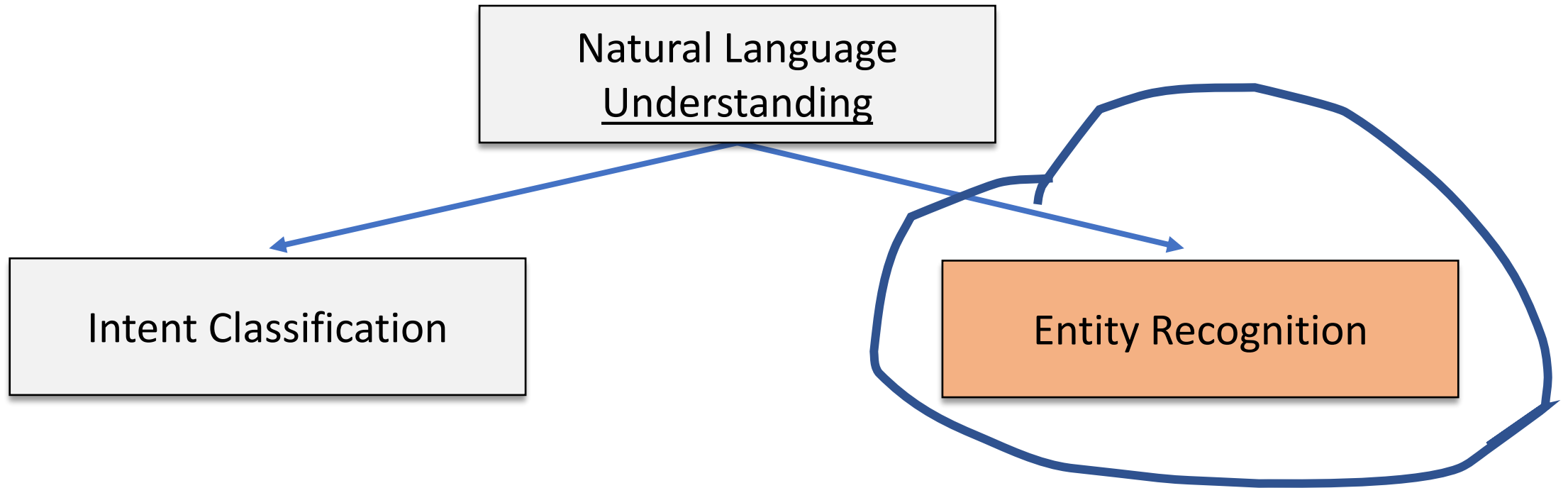
Lifelong Learning framework

MeLL: Large-scale Extensible User Intent Classification for Dialogue Systems with Meta Lifelong Learning

Chengyu Wang[1]*, Haojie Pan[1]*, Yuan Liu[1], Kehan Chen[1], Minghui Qiu[1], Wei Zhou[1], Jun Huang[1], Haiqing Chen[1], Wei Lin[1], Deng Cai[2]

[1] Alibaba Group [2] State Key Lab of CAD & CG, Zhejiang University

a) Initial Learning Stage

b) Lifelong Learning Stage

https://chywang.github.io/papers/kdd2021.pdf

Natural Language Understanding

Intent Classification

Entity Recognition

Entity recognition is the task of identifying and extracting structured information from the unstructured text.

(Named) Entity Recognition

In fact, the Chinese [NORP] market has the three [CARDINAL] most influential names of the retail and tech space – Alibaba [GPE] , Baidu [ORG] , and Tencent [PERSON] (collectively touted as BAT [ORG]), and is betting big in the global AI [GPE] in retail industry space . The three [CARDINAL] giants which are claimed to have a cut-throat competition with the U.S. [GPE] (in terms of resources and capital) are positioning themselves to become the 'future AI [PERSON] platforms'. The trio is also expanding in other Asian [NORP] countries and investing heavily in the U.S. [GPE] based AI [GPE] startups to leverage the power of AI [GPE] . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one [CARDINAL] , with an anticipated CAGR [PERSON] of 45% [PERCENT] over 2018 - 2024 [DATE] .

To further elaborate on the geographical trends, North America [LOC] has procured more than 50% [PERCENT] of the global share in 2017 [DATE] and has been leading the regional landscape of AI [GPE] in the retail market. The U.S. [GPE] has a significant credit in the regional trends with over 65% [PERCENT] of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google [ORG] , IBM [ORG] , and Microsoft [ORG] .

# Token Classification vs. Text Classification

User utterances	Intent Label
Book a flight from LA to Madrid.	Book flight
Can you get me a table at Steve's?	Restaurant reservation
Book Delta ticket Madison to Atlanta.	Book flight
Schedule me a table at Red Lobster.	Restaurant reservation
...	...
Can you tell me the name of this song?	Open
Look up the calories in an apple.	Open

Tokens are provided labels.

Token labels are predicted by machine

Text sequence is provided a label.

Hence, text label is predicted by machine

Same evaluation technique used for both tasks.

Dialogue Flow
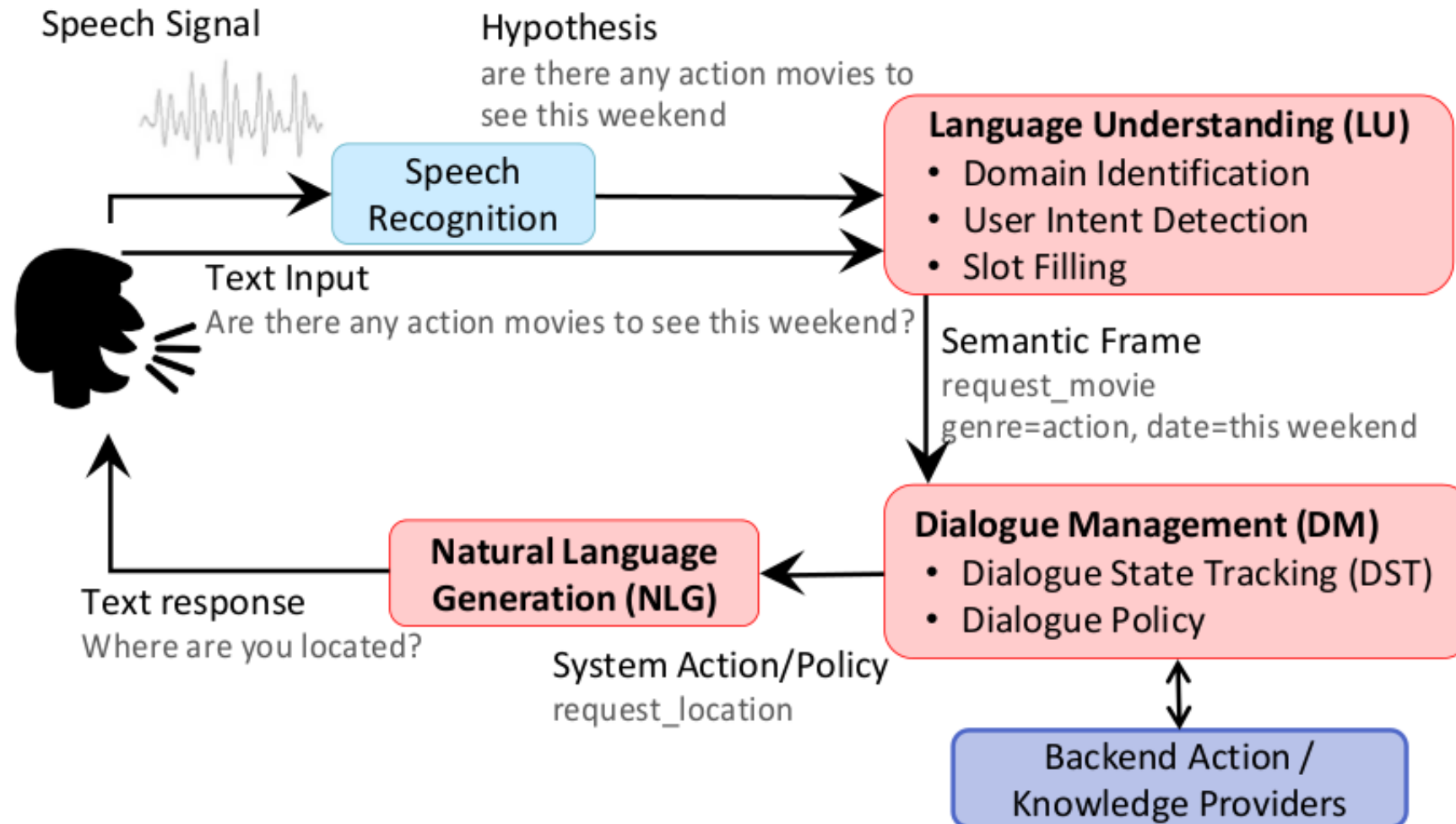
Challenges with Dialogue Flow

- Task oriented chat vs. chit-chatting.
- Change in user intent.
- Dialogue state management – continuation of a leftover conversation.

- Start with heuristics or rules – then go for generation.

- Language Generation!

Other NLP Investigations

- Abbreviation Detection (LREC 2022)
  - We contribute a large resource for detection of abbreviations and acronyms in the scientific domain.
- Named Entity Recognition (LREC 2022)
  - We contribution the largest known manually annotated Hindi NER corpus.
- Offensive Language Identification for Indic languages (ongoing for Malayalam)
- Using Abstractive Summarization to automate the fact-checking pipeline (Submitted to COLING 2022)
- Investigation on use of Cognitive features to improve NLP task performance.

and many more…

Thank you!

- Questions?