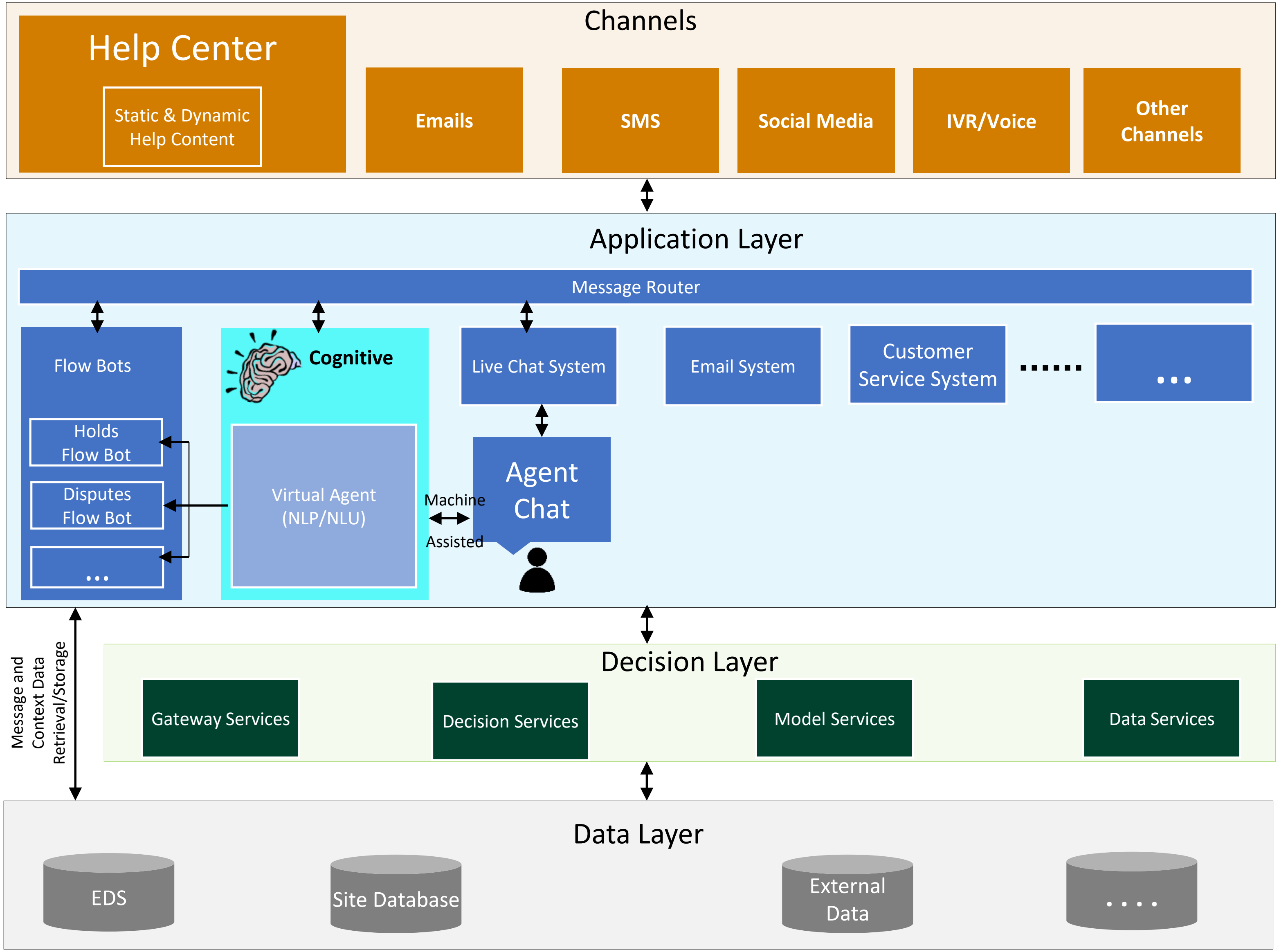
Improving Customer Service with Deep Learning Techniques in a Multi-Touchpoint System

Rajesh Munavalli
PayPal Inc

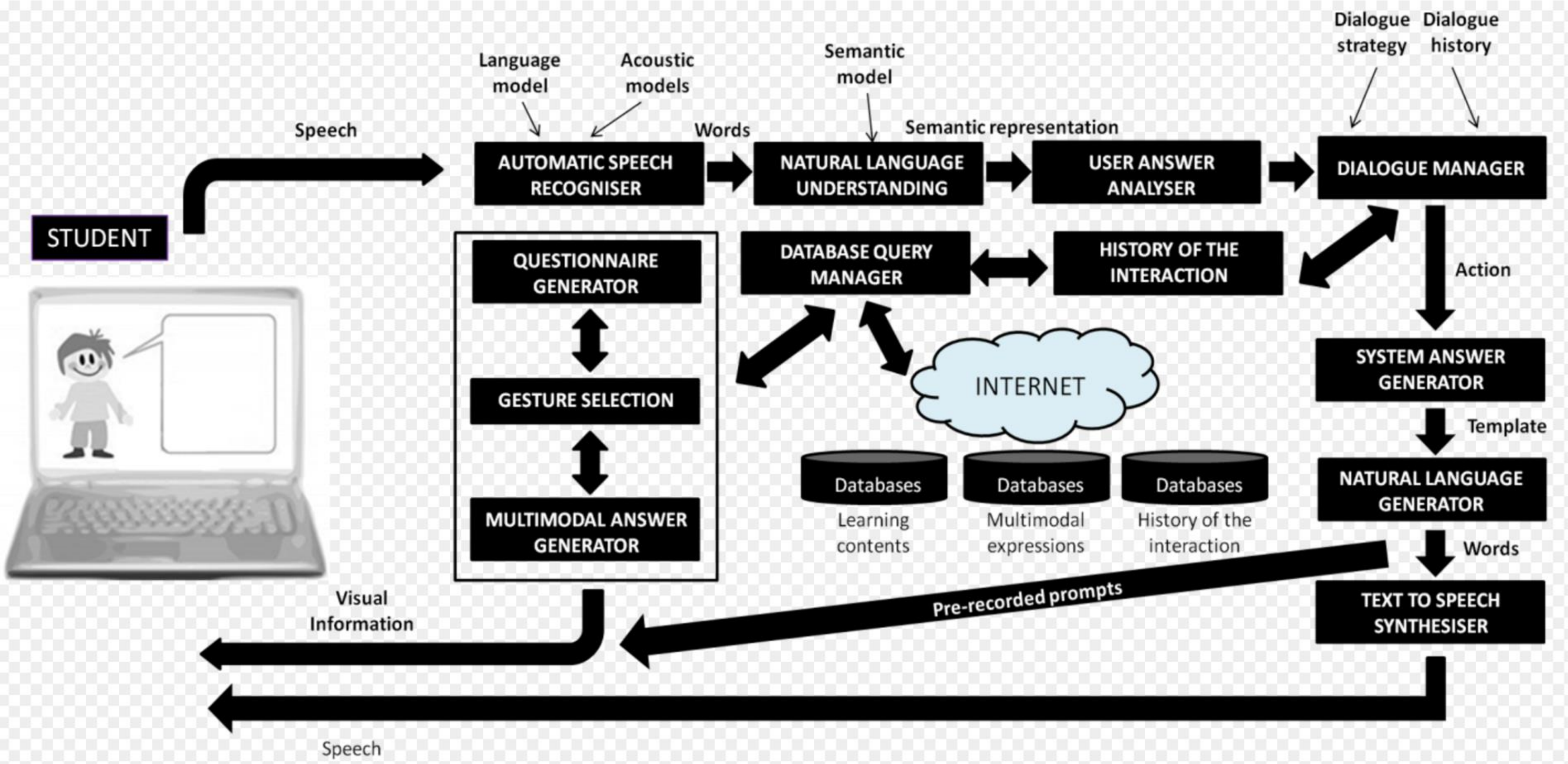
Outline

- PayPal Customer Service Architecture
- Evolution of NLP
- Help Center and Email Routing Projects
- Why Deep Learning?
- Deep Learning Architectures
  - Word Embedding
  - Unlabeled Data
- Results an Benchmarks
- Future Research

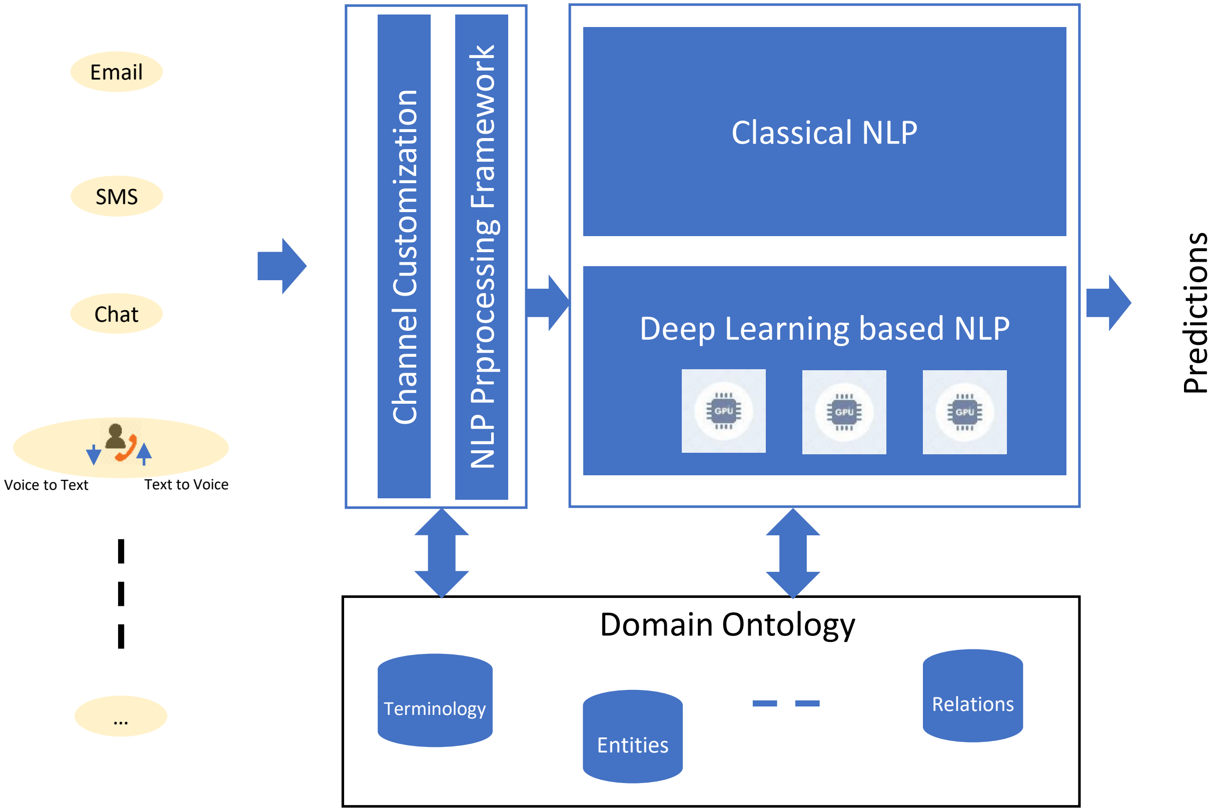
System Architecture



ChatBot Architecture



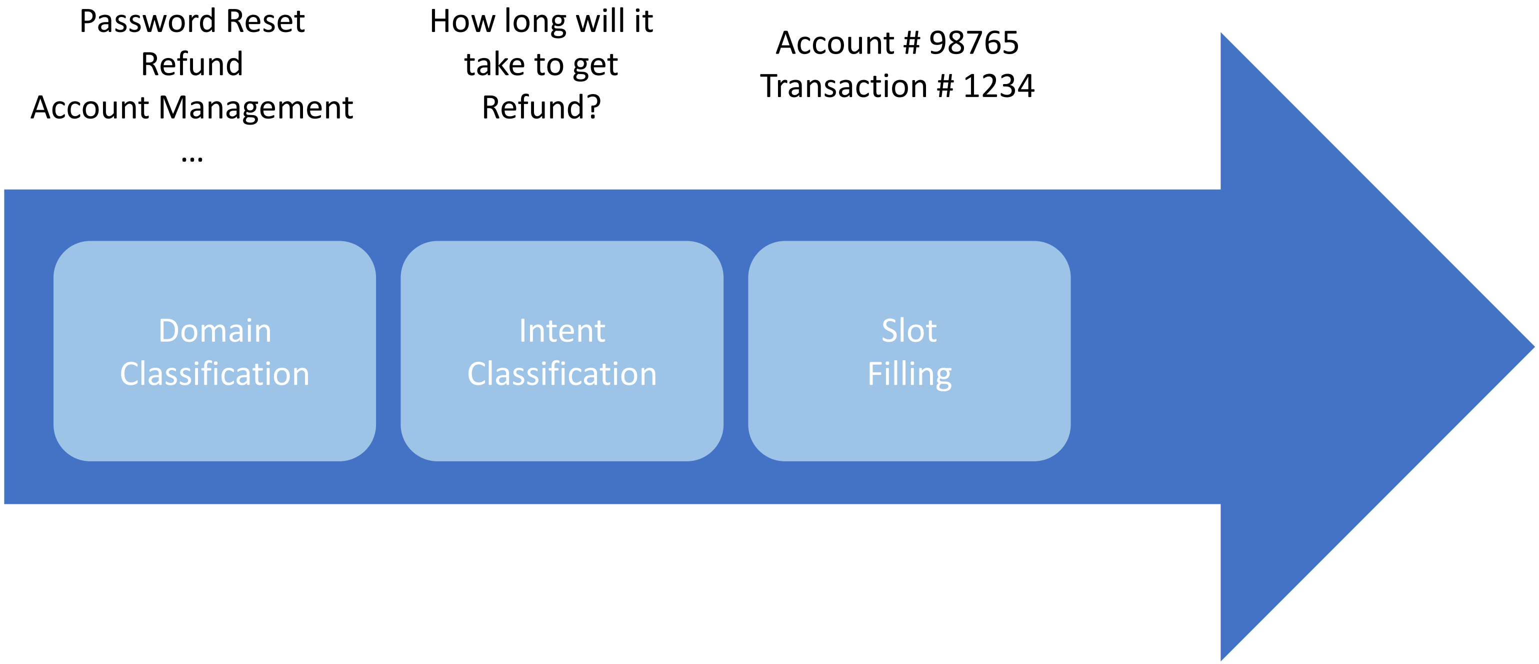
Overall NLU Architecture



Customer Service Management Core Components

- Natural Language Processing to understand user input
  - Information Extraction
  - Intent Prediction
- Dialogue and Context Management to continue conversation intelligently
- Business Logic and Intelligence
- Connectivity with the external systems to provide necessary information and take actions on behalf of the user

Information Extraction



Password Reset
Refund
Account Management
…

How long will it take to get Refund?

Account # 98765
Transaction # 1234

Domain Classification

Intent Classification

Slot Filling

Information Extraction

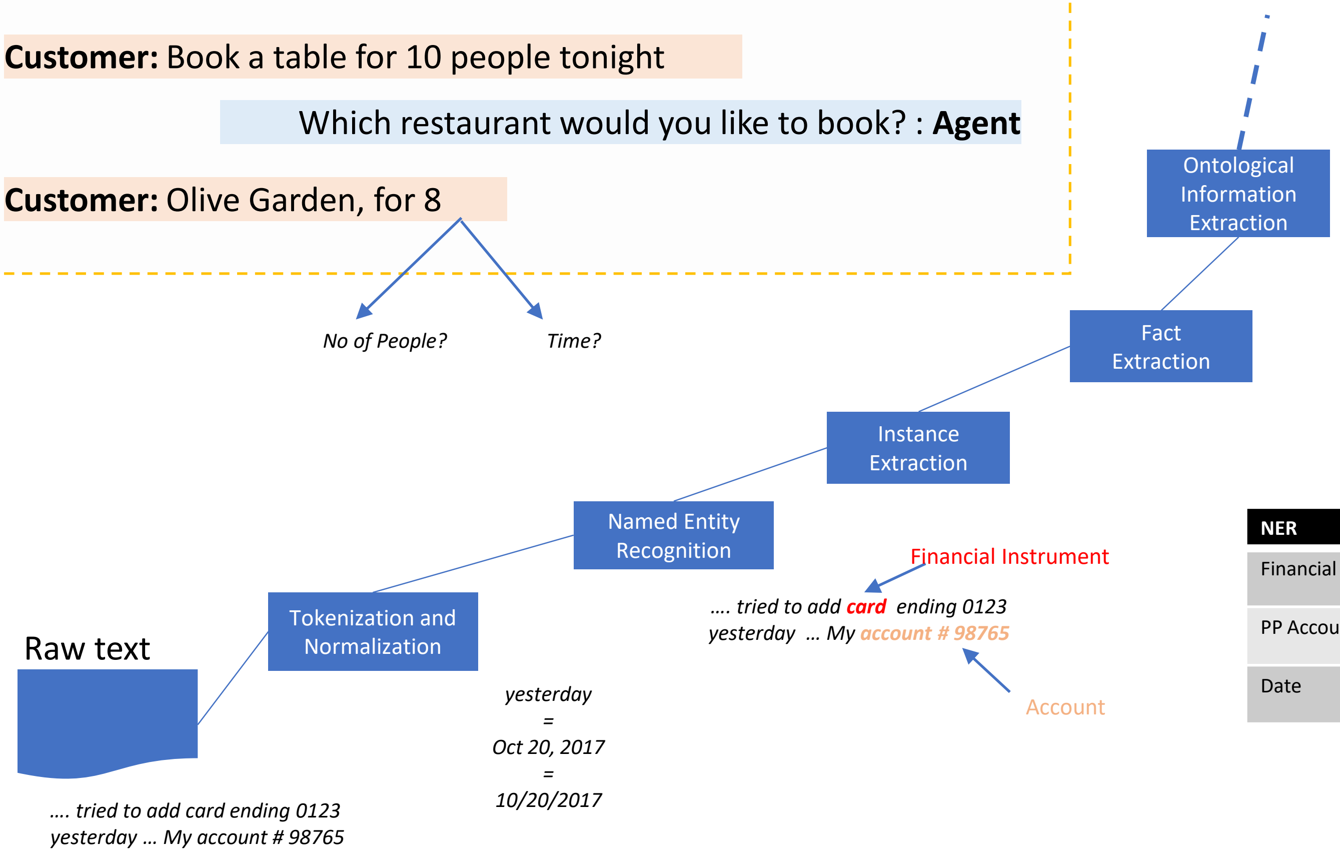
Customer: Book a table for 10 people tonight

Which restaurant would you like to book? : **Agent**

Customer: Olive Garden, for 8

No of People?

Time?



Ontological Information Extraction

Fact Extraction

Instance Extraction

Named Entity Recognition

Tokenization and Normalization

Raw text

…. tried to add card ending 0123 yesterday … My account # 98765

yesterday
=
Oct 20, 2017
=
10/20/2017

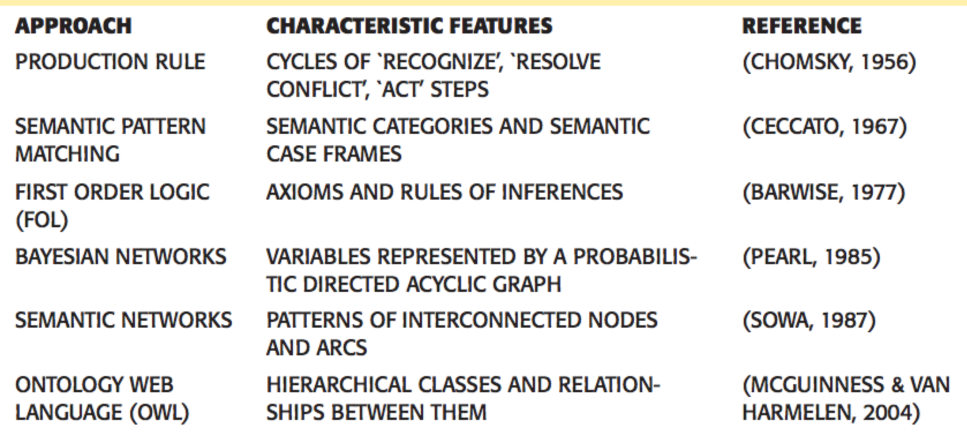
Financial Instrument

*…. tried to add **card** ending 0123 yesterday … My **account # 98765***

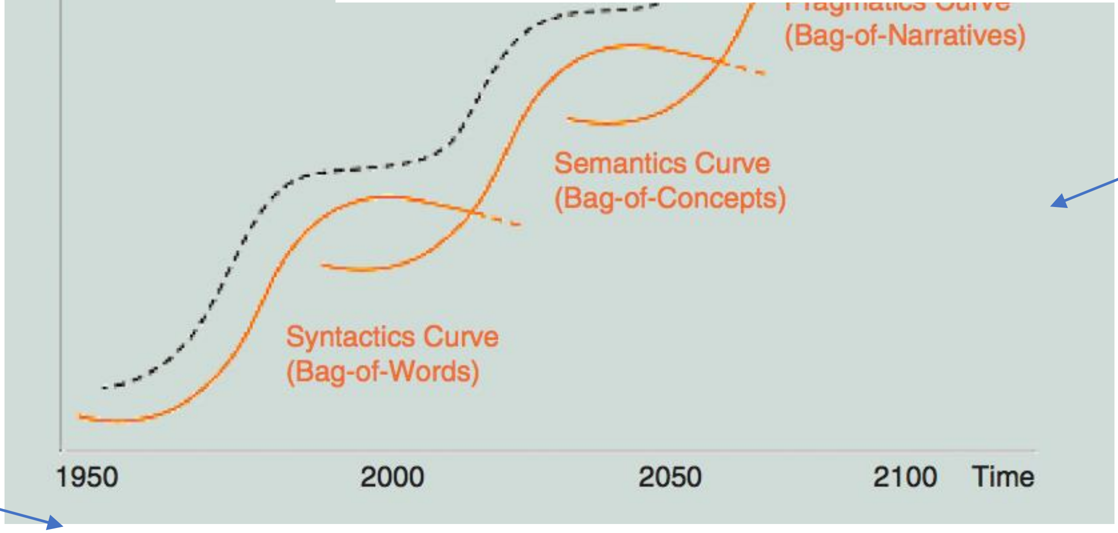
Account

NER	Instance
Financial Instrument	Card ending 0123
PP Account	98765
Date	10/20/2017

Evolution of NLP/NLU



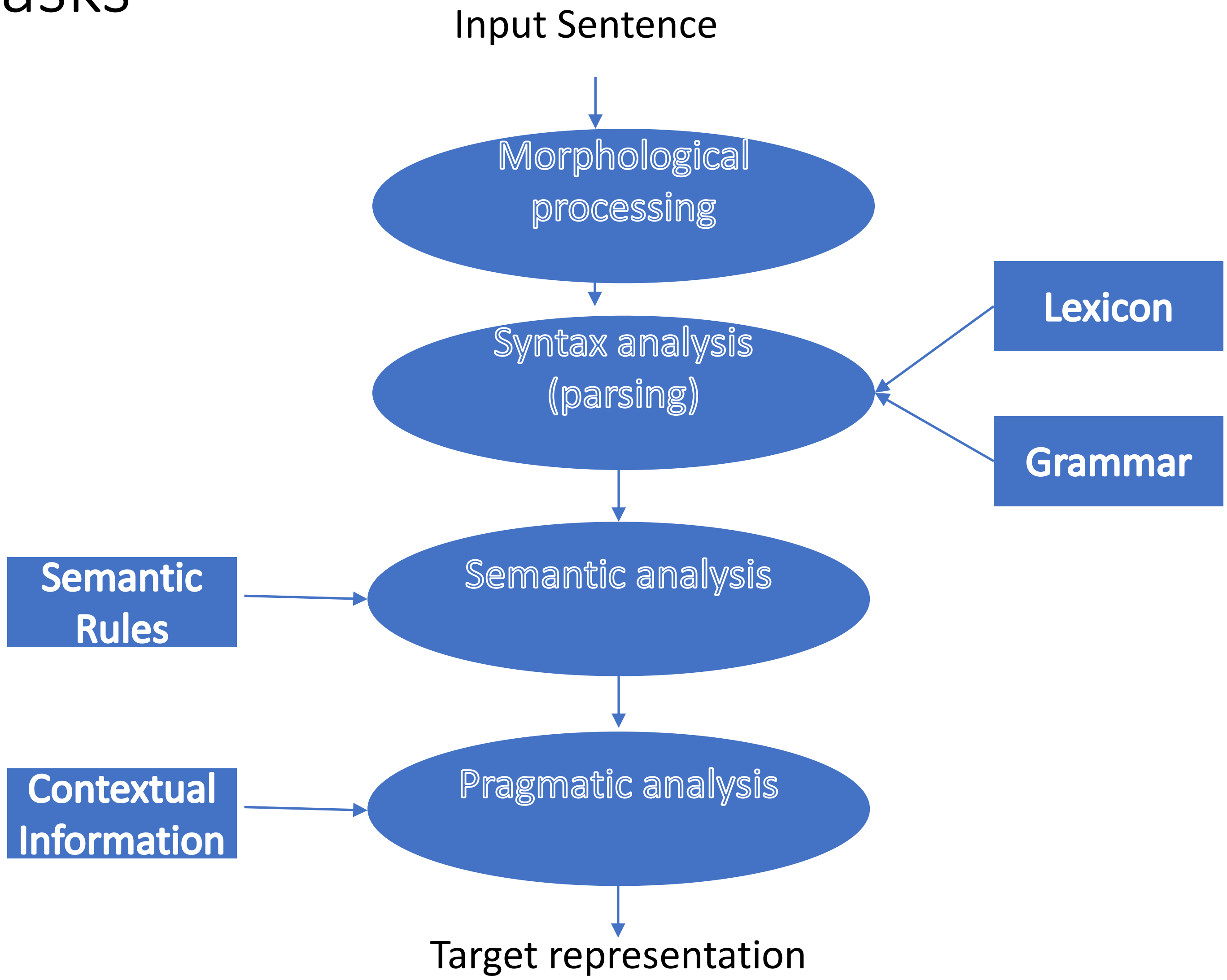
APPROACH	CHARACTERISTIC FEATURES	REFERENCE
PRODUCTION RULE	CYCLES OF 'RECOGNIZE', 'RESOLVE CONFLICT', 'ACT' STEPS	(CHOMSKY, 1956)
SEMANTIC PATTERN MATCHING	SEMANTIC CATEGORIES AND SEMANTIC CASE FRAMES	(CECCATO, 1967)
FIRST ORDER LOGIC (FOL)	AXIOMS AND RULES OF INFERENCES	(BARWISE, 1977)
BAYESIAN NETWORKS	VARIABLES REPRESENTED BY A PROBABILISTIC DIRECTED ACYCLIC GRAPH	(PEARL, 1985)
SEMANTIC NETWORKS	PATTERNS OF INTERCONNECTED NODES AND ARCS	(SOWA, 1987)
ONTOLOGY WEB LANGUAGE (OWL)	HIERARCHICAL CLASSES AND RELATIONSHIPS BETWEEN THEM	(MCGUINNESS & VAN HARMELEN, 2004)



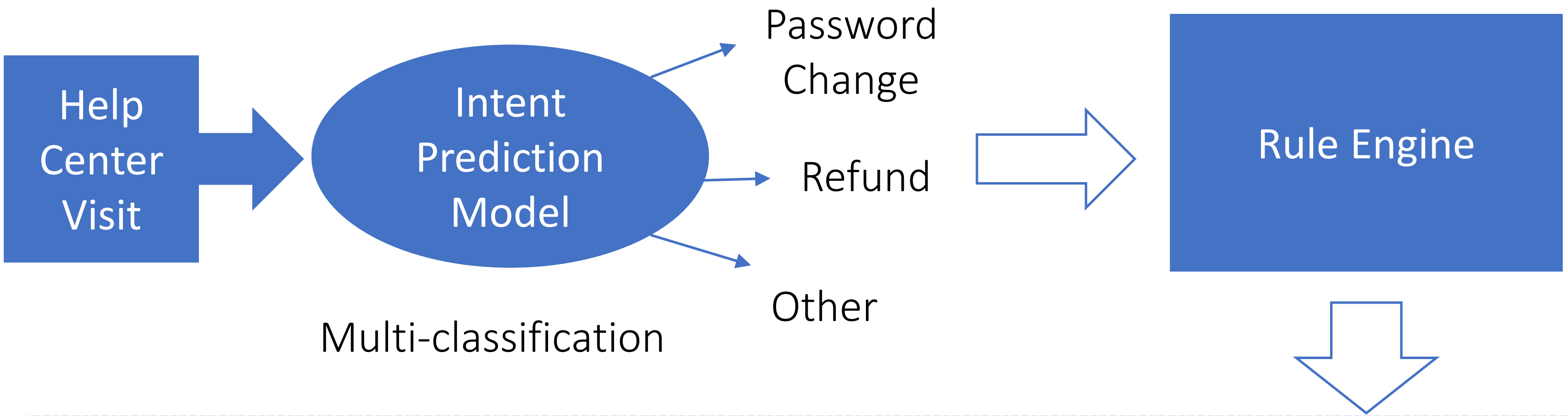
NLU

NLP

NLP Tasks


Input Sentence
Morphological processing
Syntax analysis (parsing)
Lexicon
Grammar
Semantic Rules
Semantic analysis
Contextual Information
Pragmatic analysis
Target representation

Help Center:
Intent Prediction Solution Architecture


Help Center Visit

Intent Prediction Model

Multi-classification

Password Change

Refund

Other

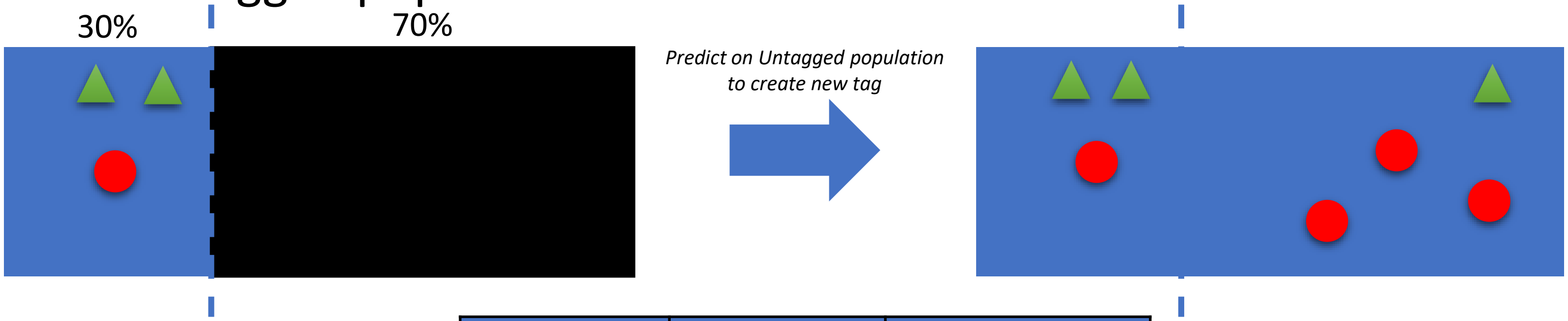
Rule Engine

BNA Use Case	Rank high likelihood intent as #1 on FAQ
Channel Steering Use Case	Pre-populate high likelihood intent on ‘Contact Us’ page

Where do we get the tags?

Iterative learning to fill gap between tagged and untagged population

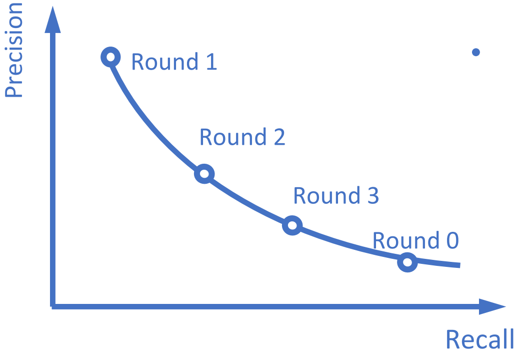
- We use the tagged population to identify “look alike” population in the untagged population



Predict on Untagged population to create new tag

Iterative Learn	Distribution	%change from base
Others	75.4%	-3%
GETMONEYBACK	8.2%	2%
PAYREF001	5.0%	20%
PAYDEC001	3.5%	6%
DISPSTATUS001	3.2%	21%
PAYHOLD001	2.9%	30%
DISPLIM001	1.9%	7%

Iterative learning boosts precision overall from 65% baseline to 79%



- Iterative learning is an optimization between precision and recall.

	Training Data	Precision on Tagged Population	Recall on Tagged Population	Manual Review Precision on tagged + untagged population	Manual Review Precision on untagged population
Round 0 (Baseline)	Tagged population	51%	69%	65%	45%
Round 1	Tagged population + untagged population as 'Other'	81%	29%	81%	68%
Round 2	Tagged population + round 1 prediction for untagged population	77%	33%	79%	70%
Round 3	Tagged population + round 2 prediction for untagged population	75%	36%	76%	67%

Taxonomy of Models

- Retrieval based vs Generative based
  - Retrieval (Easier):
    - No new text is generated
    - Repository of predefined responses with some heuristic to pick the best response
    - Heuristic could be as simple as rule-based expression or as complex as ensemble of classifiers
    - Wont be able to handle unseen cases and context
  - Generative (Harder):
    - Generate new text
    - Based on MT Techniques but generalized to input sequence to output sequence
    - Quite likely to make grammatical mistakes but smarter

Challenges

- **Short vs Long Conversations**
  - Shorter conversations (Easier)
    - Easier and goal is usually to create single response to a single input
    - Ex: Specific question resulting in a very specific answer
  - Longer conversations (Harder)
    - Harder and often ambiguous on the intent of the user
    - Need to keep track of what has been already said and sometimes need to forget what has been already discussed

Closed vs Open Domain:

- Closed Domain (Easier):
  - Most of the customer support systems fall into this criteria
  - How do we handle new use case? Product?
- Open Domain (Harder):
  - Not relevant to our use cases

Challenges

- **Incorporating Context**
  - Longer conversations (Harder)
    - Harder and often ambiguous on the intent of the user
    - Need to keep track of what has been already said and sometimes need to forget what has been already discussed

Coherent Personality

- Closed Domain (Easier):
  - Most of the customer support systems fall into this criteria

Evaluation of models

- Subjective
- BLEU score – Extensively used in MT systems

Intention and Diversity

- Most common problem with Generative models is providing a generic canned response like “Great”, “I don’t know”..etc

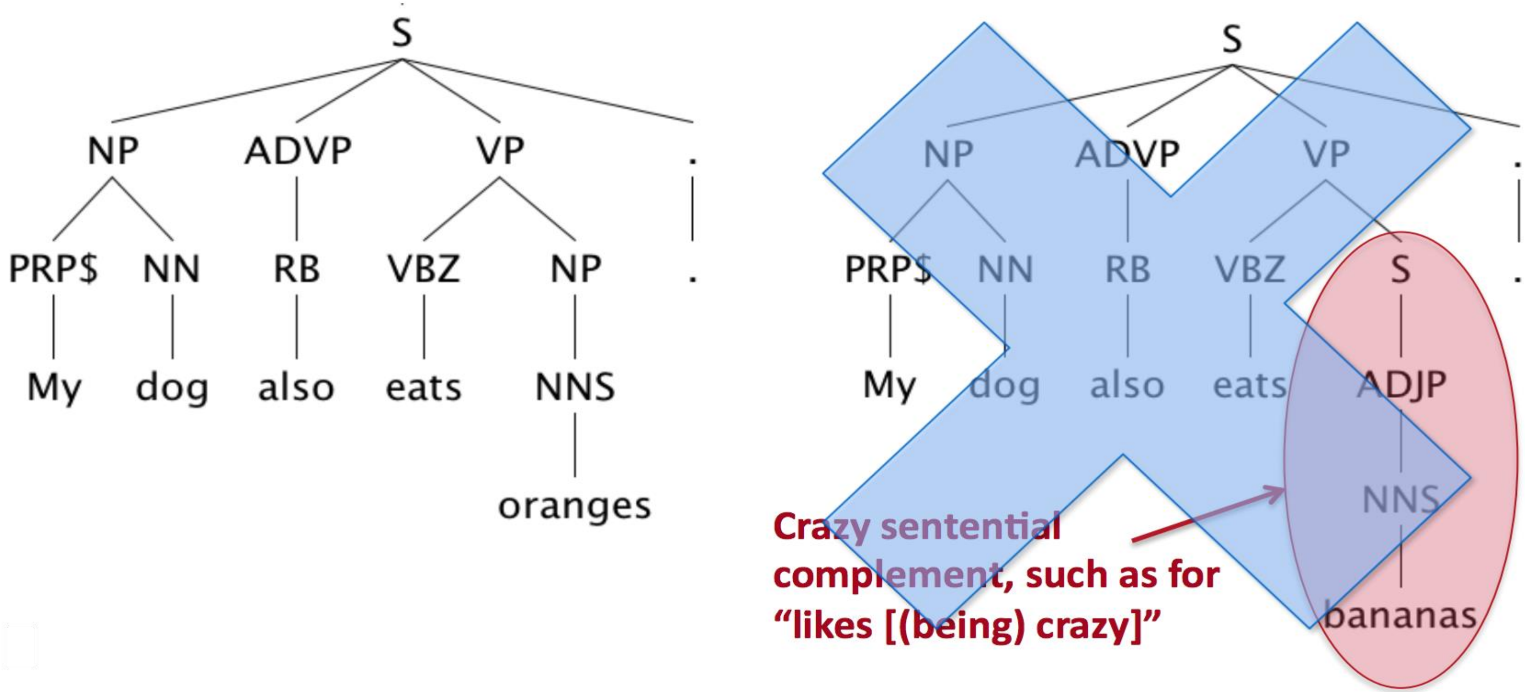
Why Deep Learning?

Automatic learning of features

- Traditional Feature Engineering
  - Time Consuming
  - Most of the time over-specified (repetitive)
  - Incomplete and not-exhaustive
  - Domain Specific and needs to be repeated for other domains

Why Deep Learning?

Generalized/Distributed Representations



- Distributed representations help NLP by representing more dimensions of similarity
    - Tackles Curse of dimensionality

Why Deep Learning?

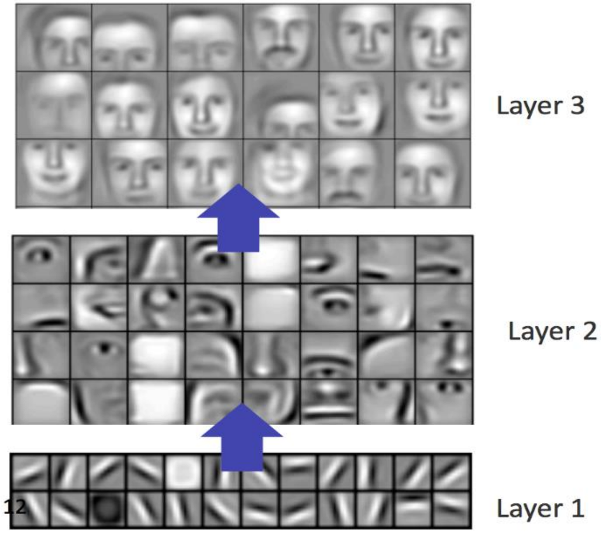
Unsupervised feature and weight learning

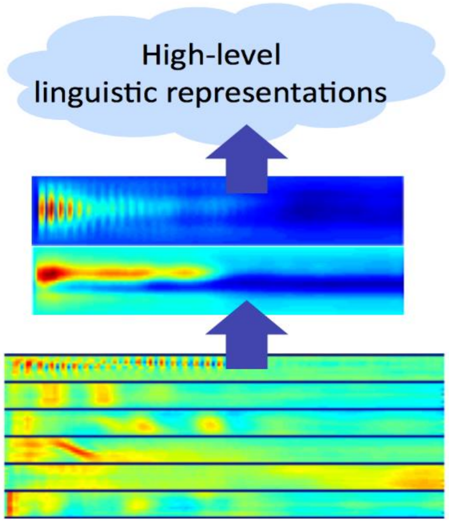
- Almost all good NLP & ML methods need labeled data. But in reality most data is unlabeled

- Most information must be acquired unsupervised

Why Deep Learning?
Hierarchical Feature Representation

- Hierarchical feature representation
  - Biologically inspired
  - Brain has deep architecture
  - Need good intermediate representations shared across tasks
  - Human language is inherently recursive



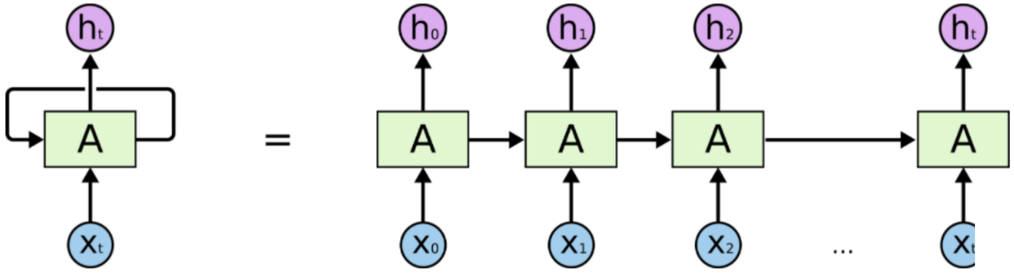


Why Deep Learning?
Why now?

Why methods failed prior to 2006?

- Efficient parameter estimation methods
- Better understanding of model regularization
- New methods for unsupervised training: RBMs (Restricted Boltzmann Machines), Autoencoders..etc

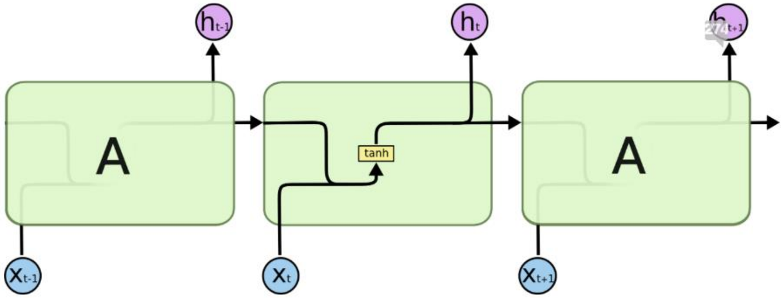
RNNs



RNN Concept

Unrolled RNN equivalent

Repeating module in a standard RNN contains a single layer



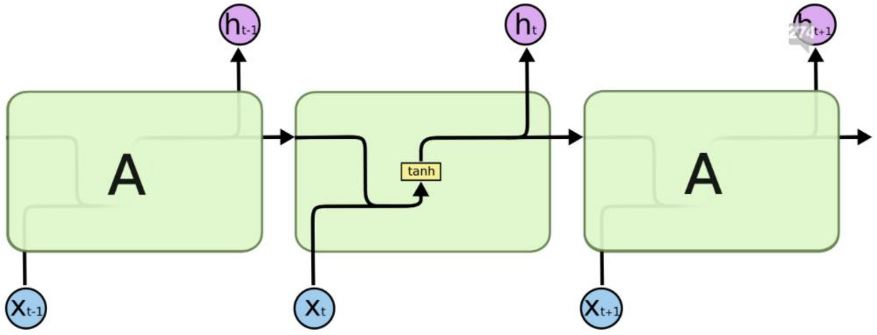
Context Matters

Tackle with Distributed similarity

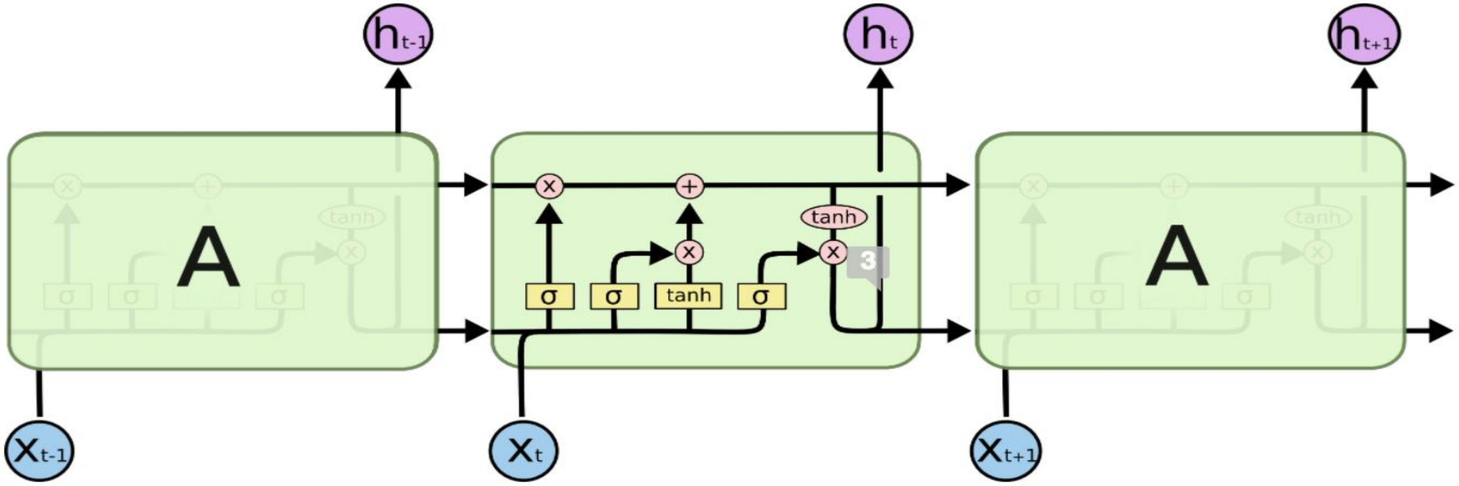
CFPB today sued the River Bank over consumer allegations

We walked along the river bank

LSTMs and GRUs

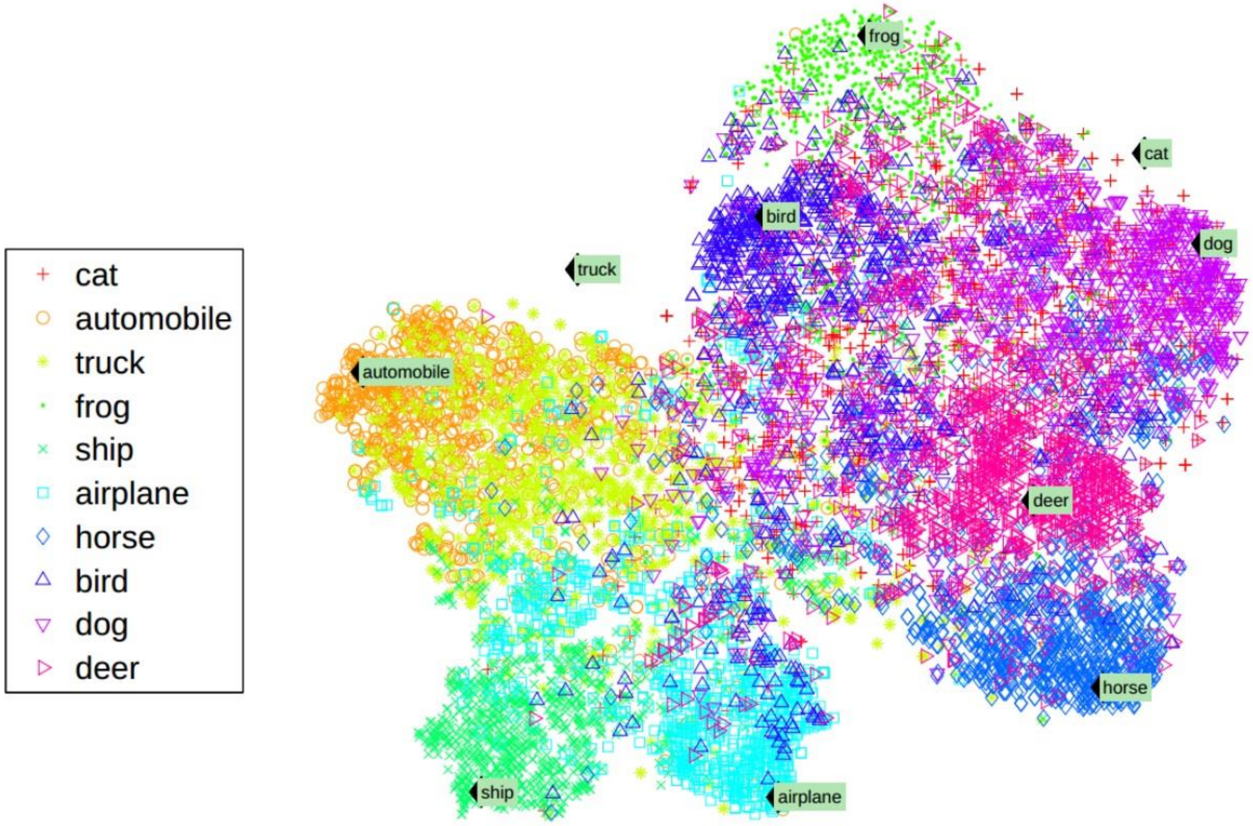


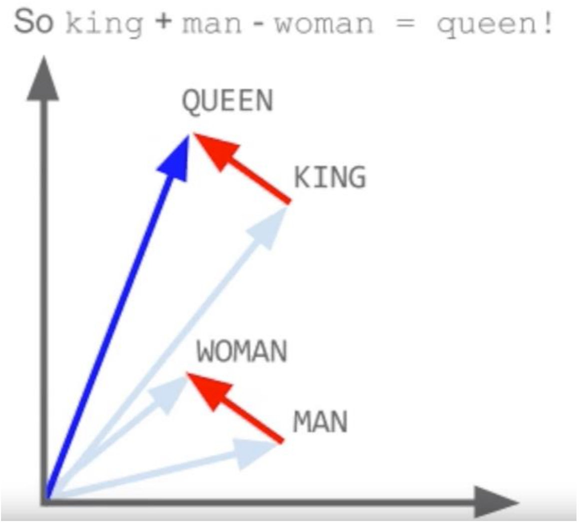
Repeating module in a standard RNN contains a single layer



LSTM repeating module has 4 interacting layers

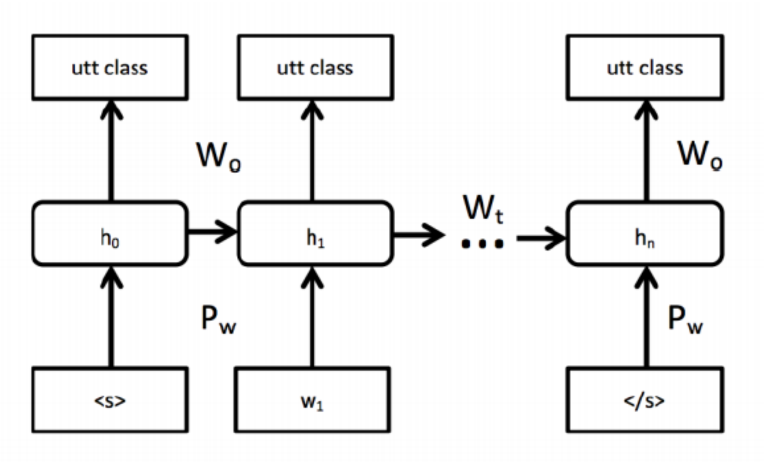
Leveraging Unlabeled Data
Word Embedding - Word2Vec

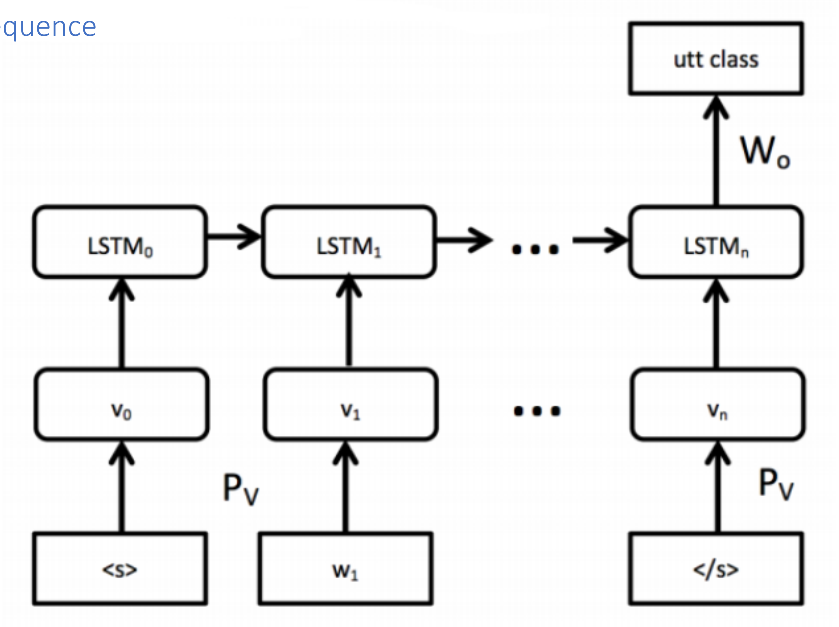



So king + man - woman = queen!

Domain/Intent Classification

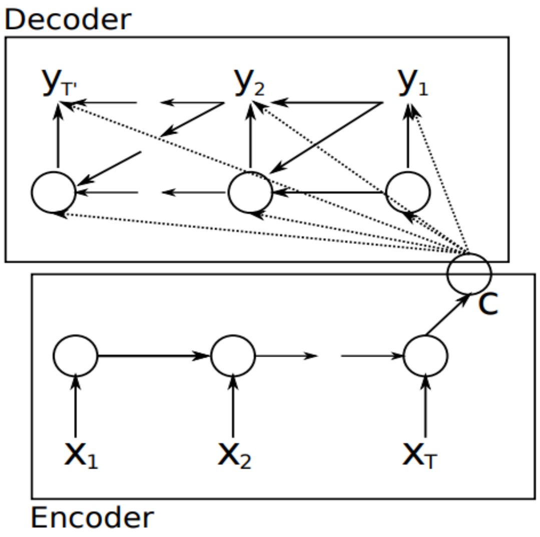
- Sequences can be either a single chat message or an entire email
- Intent classification performs better when applied to the entire sequence

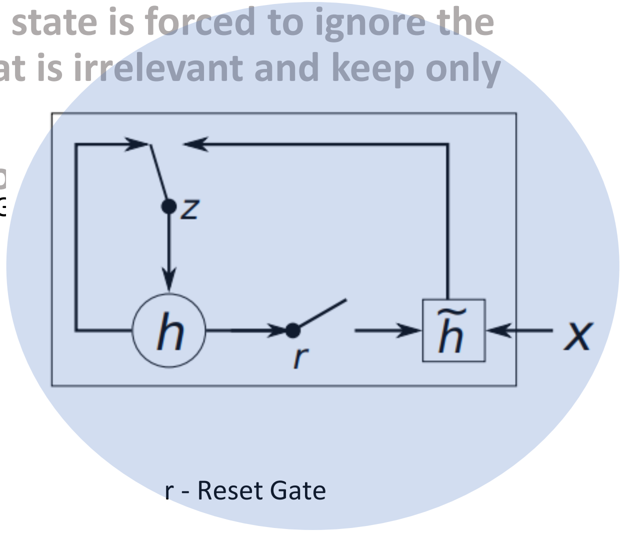




Example: Sequence to Sequence Modeling

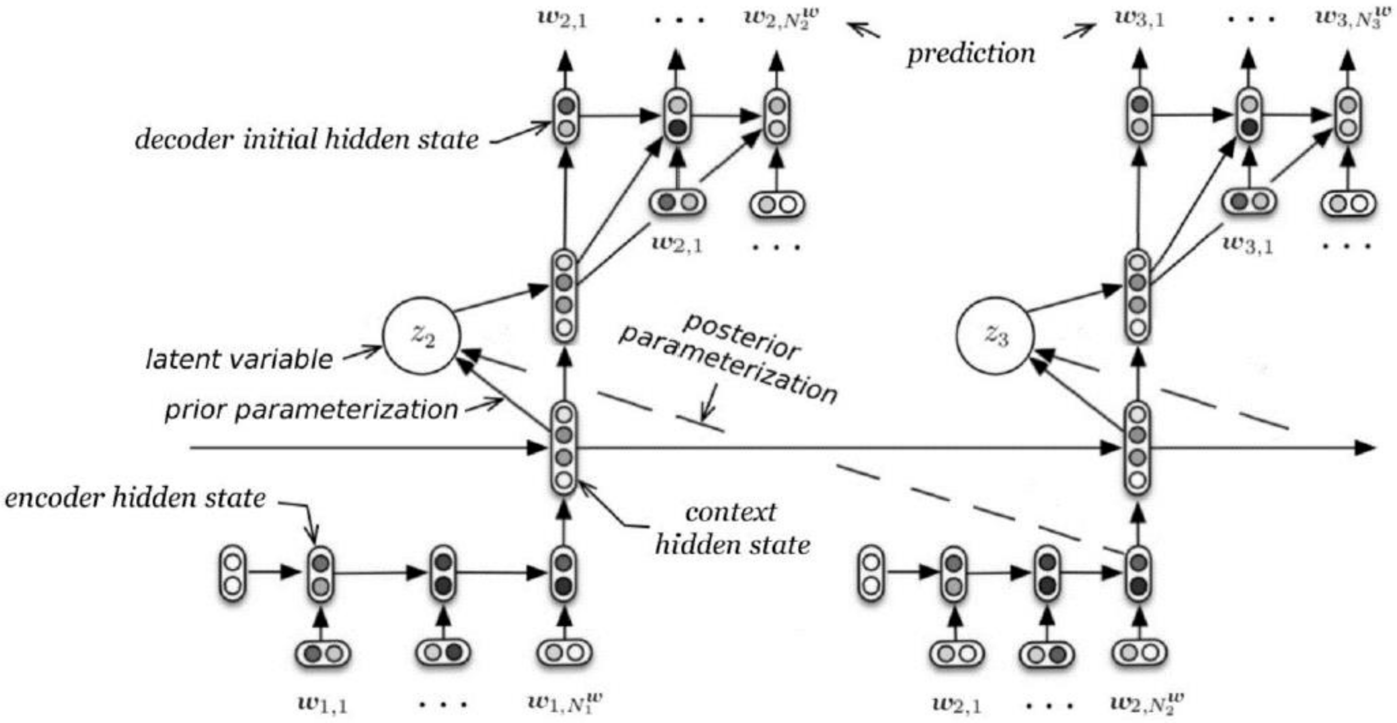
- **Learns to encode a variable length sequence into a fixed length vector representation**
- **Decode a given fixed-length vector representation back into a variable length sequence**
- **Gate functionality**
  - **R (short term) - when reset gate is close to 0, the hidden state is forced to ignore the previous hidden state thus dropping any information that is irrelevant and keep only the current**
  - **Z (long ter ormation fro over acting**





Hidden Activation function

End-to-End Deep Learning


$w_{2,1}$ . . . w_{2,N_2^w}
prediction
$w_{3,1}$ . . . w_{3,N_3^w}
decoder initial hidden state
$w_{2,1}$ . . .
$w_{3,1}$ . . .
z_2
z_3
latent variable
posterior parameterization
prior parameterization
encoder hidden state
context hidden state
$w_{1,1}$ . . . w_{1,N_1^w}
$w_{2,1}$ . . . w_{2,N_2^w}

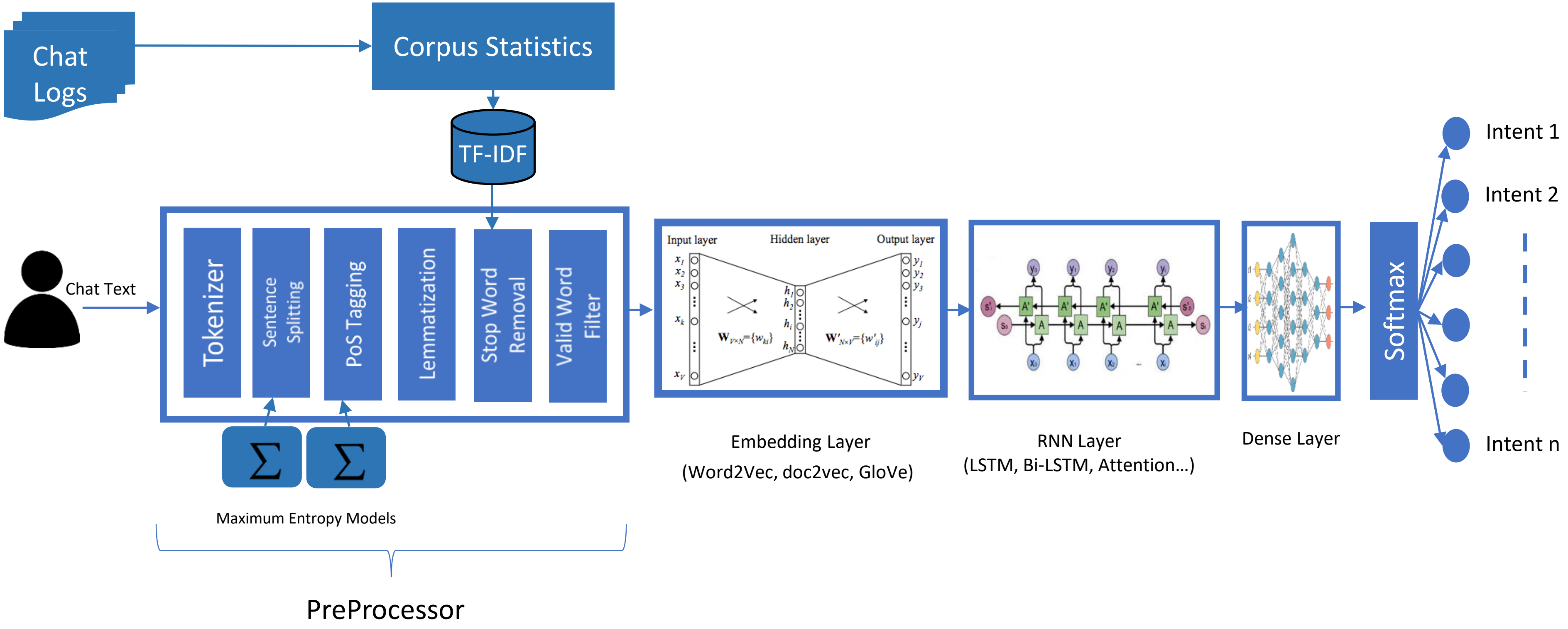
Which transaction?

Transaction #1234

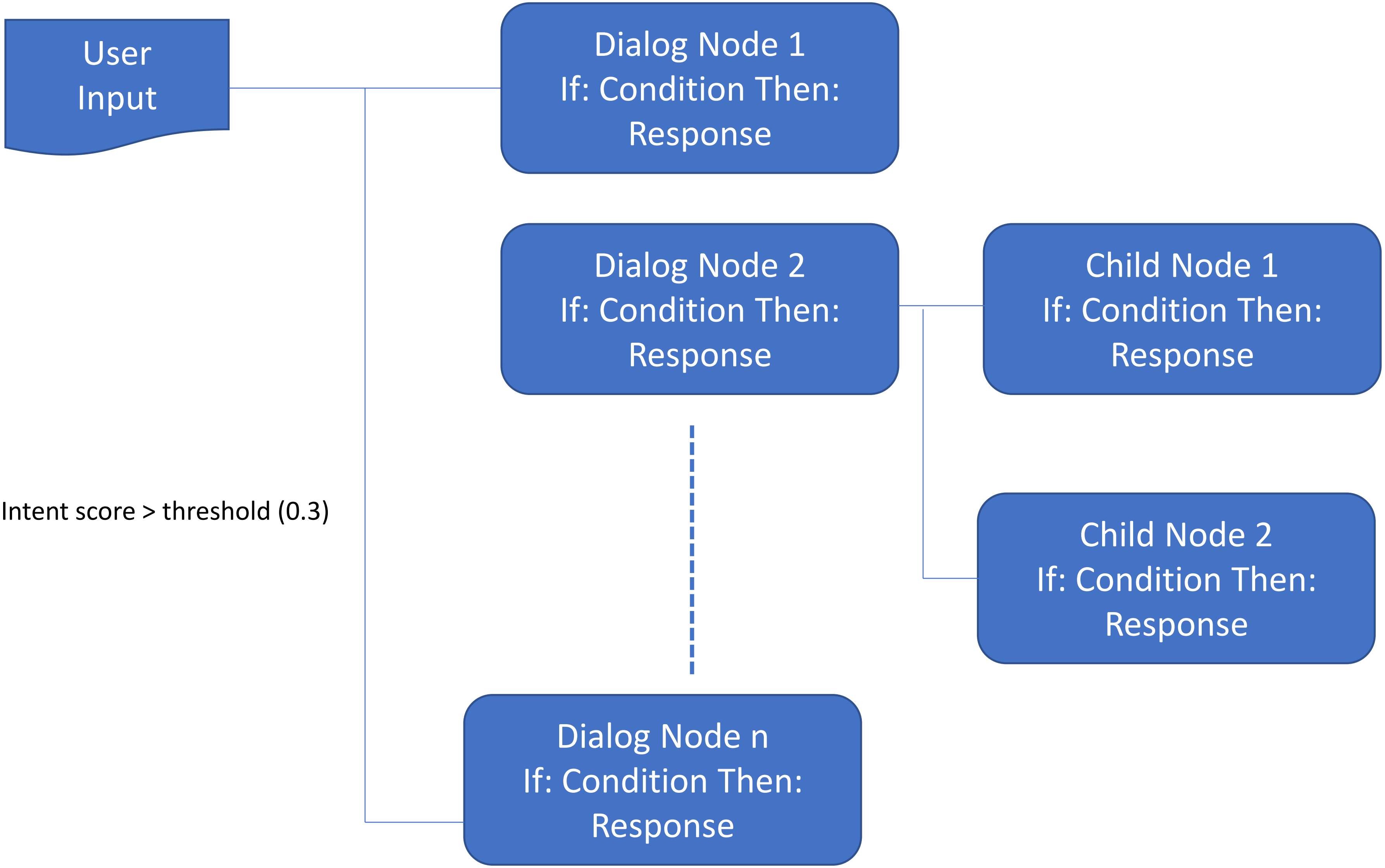
When would I get refund?

Which transaction?

Intent Prediction Model



Dialog Management



User Input

Dialog Node 1
If: Condition Then: Response

Dialog Node 2
If: Condition Then: Response

Child Node 1
If: Condition Then: Response

Child Node 2
If: Condition Then: Response

Intent score > threshold (0.3)

Dialog Node n
If: Condition Then: Response

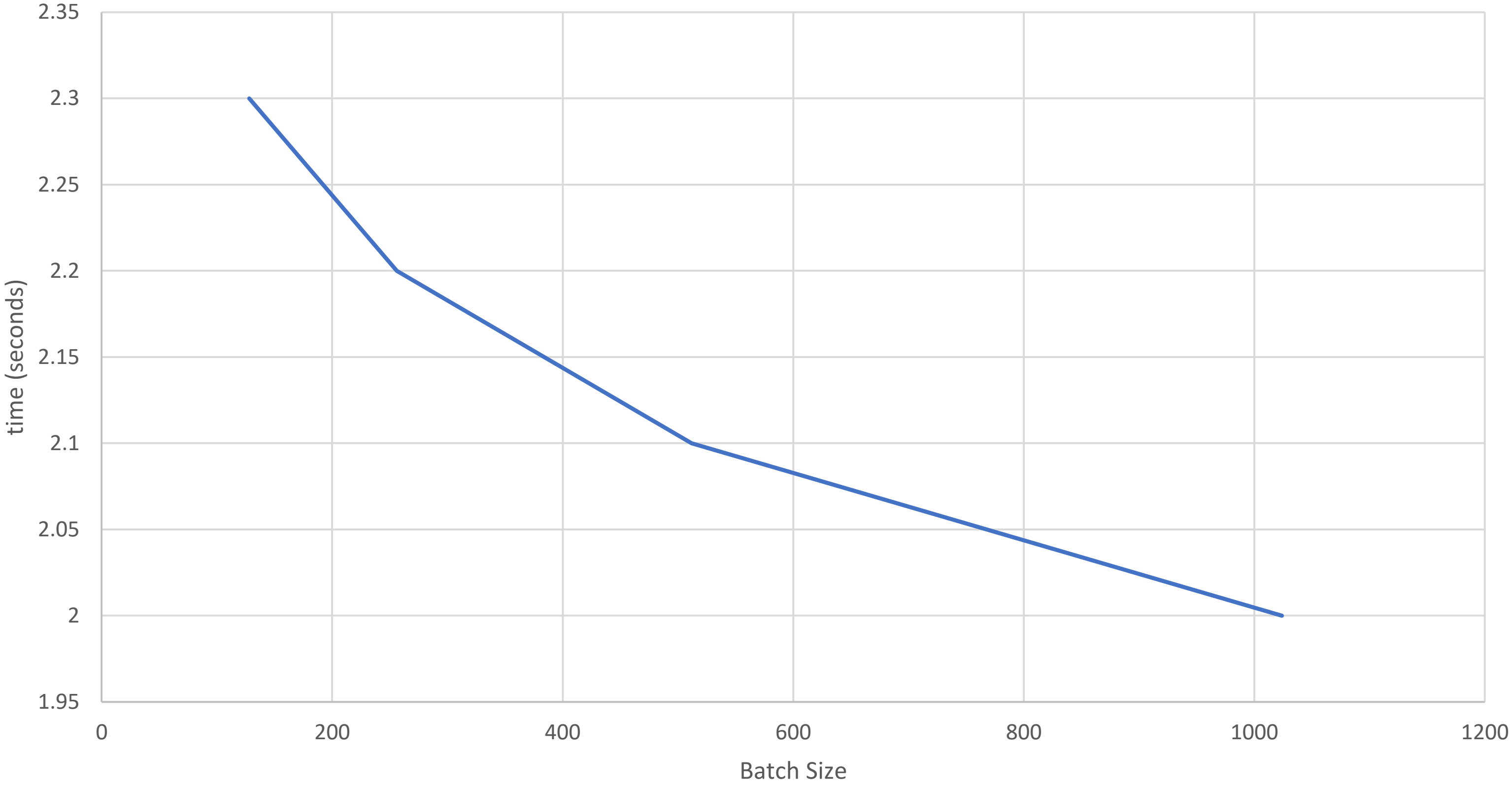
Results and Benchmarking

(NVIDIA DGX V100)

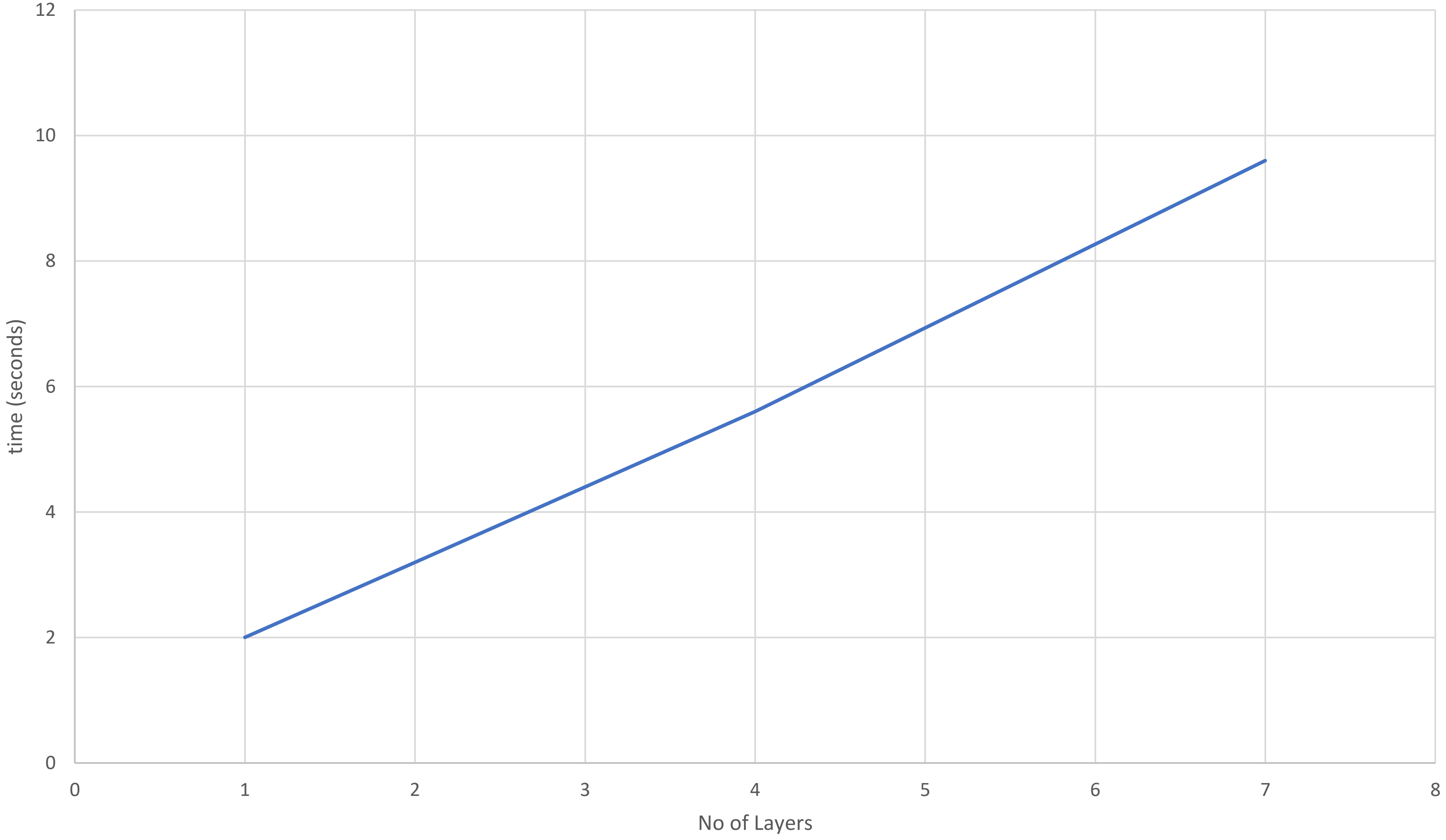
PayPal Bot vs IBM Watson

Intent	IBM Watson	LSTM	LSTM with Attention Network	Bi-Directional LSTM	Bi-Directional LSTM with Attention Network
Ask for an Agent	80.82%	91.80%	91.80%	92.50%	93.20%
End of Chat	27.27%	18.20%	9.10%	9.10%	0.00%
Greetings	88.10%	90.50%	90.50%	90.50%	90.50%
Negative Feedback	32.69%	28.80%	26.90%	32.70%	23.10%
Other	50.55%	57.10%	62.60%	62.10%	56.60%
Positive Feedback	57.14%	14.30%	28.60%	28.60%	14.30%
Refund Status	74.92%	86.10%	86.50%	84.80%	81.80%
Thank You	60.00%	90.00%	90.00%	90.00%	90.00%
Transaction/Account Details	48.68%	46.10%	40.80%	47.40%	47.40%
Overall	**65.19%**	**71.90%**	**72.70%**	**73.00%**	**70.10%**

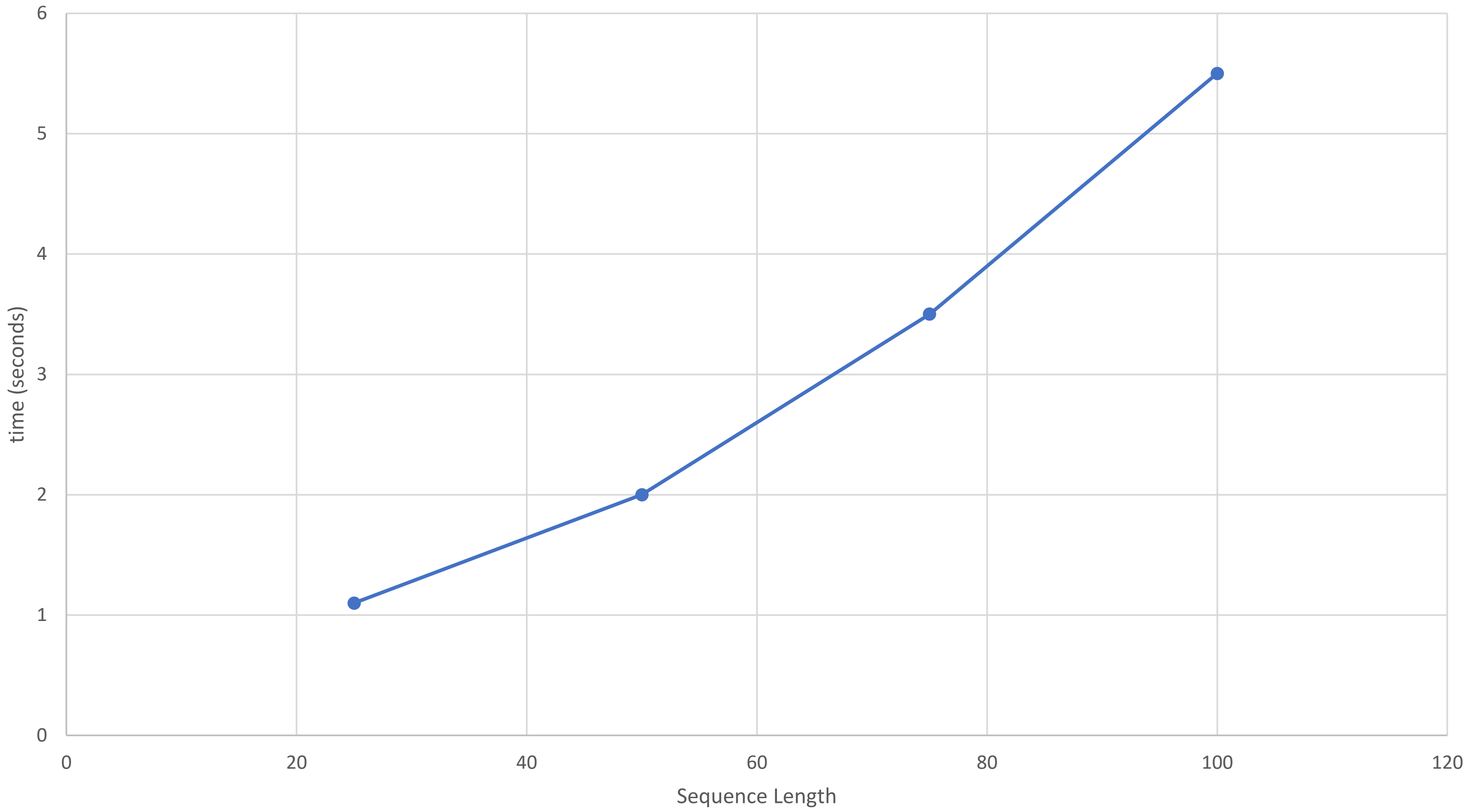
Effect of Batch Size



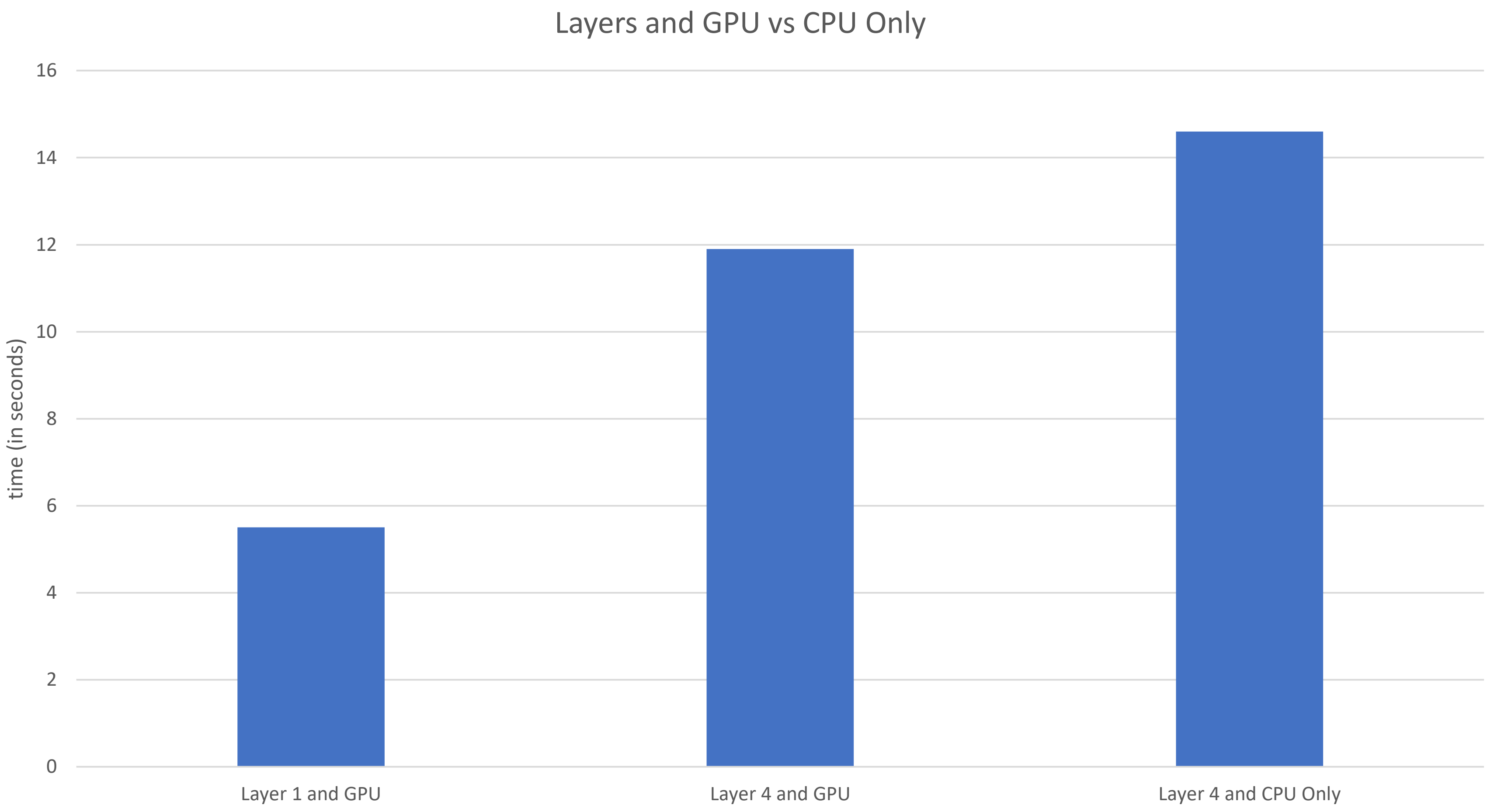
Effect of No of Layers



Effect of Sequence Length



Effect of Layers, CPU vs GPU



Future Research

- Unlabeled data augmentation
- Zero Shot/One Shot/Few Shot Learning
- Sequence to Sequence Modeling
- Averting Social Engineering/Fraud