Wednesday: NLP 05-11-24

Syntact Parsing

→ Morphology deals with internal structure of words.

→ Syntax deals with combination of words. (phares and sentences).

→ Morphology is irregular and syntax is mostly regular. → (made of rules)

→ Syntax is neither morphology nor semantics.

→ Semantics ⇒ Meaning

Syntax ⇒ Structure

Constituency → part of sentence

Phrases:

Noun Phrase

The elephant arrived.
NP

It arrived.

Elephants arrived.

The elephant I love to hate arrived.
NP

Prepositional Phrases (PP)

. I arrived on Tuesday

→ Every PP contains a noun phra

. I arrived under the leaking roof

Sentences or Clauses

→ Clause has multiple constituents

→ baat ko further explain krna clauses main aata.

⇒ It is patently false that sometimes John thinks He is Mery. (3 Clauses).

Context Free Grammars

→ Vocabulary of terminal Symbol, Σ

→ Set of non-terminal, S

Examples:

S → NP VP

NP → Det Noun

VP → Verb NP

Det → the, a

Noun → boy, girl, hotdogs

Verb → likes, hates, eats

→ girl hates hotdogs ⇒ X Incorrect
The girl Bikes

Pharsing ↔ Pharse Tree

Building Noun Phrases

NP → Determiner NounBar

NP → Proper Noun

NounBar → Noun

NounBar → AP NounBar

NounBar → NounBar PP

AP → Adj AP

AP → Adj

PP → Preposition NP

Ambiguity

→ aik part sy multiple productions

Language Myths

Subject → Syntactic

Book that Flight Top-Down Approach

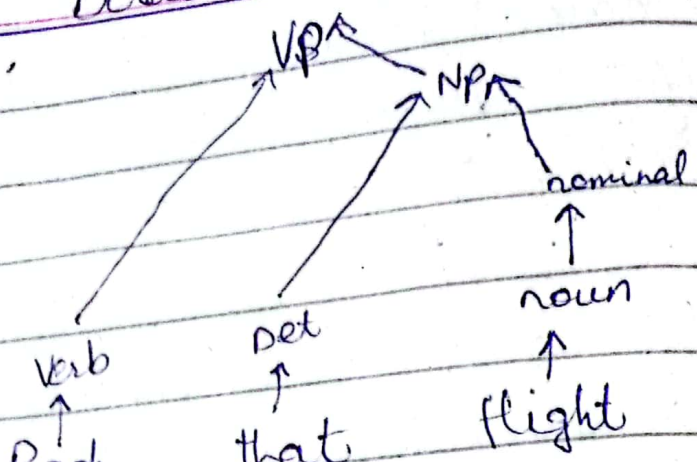
S → ~~NP~~ ⇒ VP

→ Verb NP

Book Det Noun

that → Noun → flight

Bottom - up Parser



→ Shift - redice parser.

nltk.app.rdparser() → Top-down

nltk.app.srparser() → bottom-up

Penn Treebank dekh kr aana

Friday 08-11-24

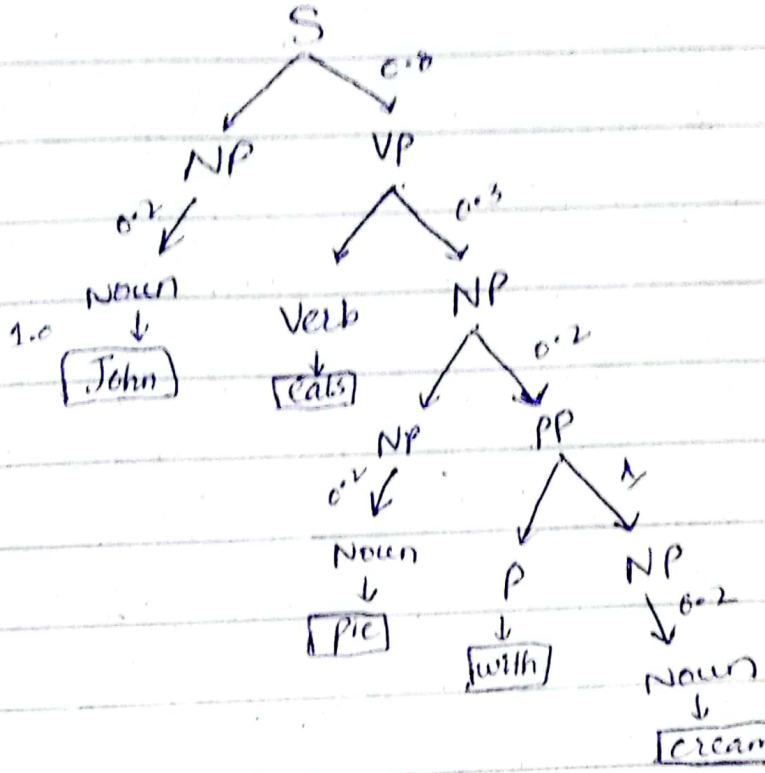
Bayes Rule

$$\text{argmax } P(\tau, S) = \text{argmax } \frac{P(\tau, S)}{P(S)}$$

Probabilistic CFG

→ Prob. of a tree τ is the product of the prob. of all its rules.

John eats pie with cream.



0.00384

$0.2 * 0.8 * 0.2 * 0.3 * 0.2 * 0.2 * 1 = 0.00384$

→ In shift reduce parser, CFG should be in CNF.

CFG to CNF

$X \rightarrow ABC \rightarrow$ CFG

$X_2 \rightarrow AX_2$] CNF

$X_2 \rightarrow BC$

Convert CFG to CNF

VP → VBD NP PP PP

Conversion:

VP → X_2 X_3

X_2 → VBD NP

X_3 → PP PP

→ Chatbot Chapter

→ Grassy

→ ht track crawler Fall back

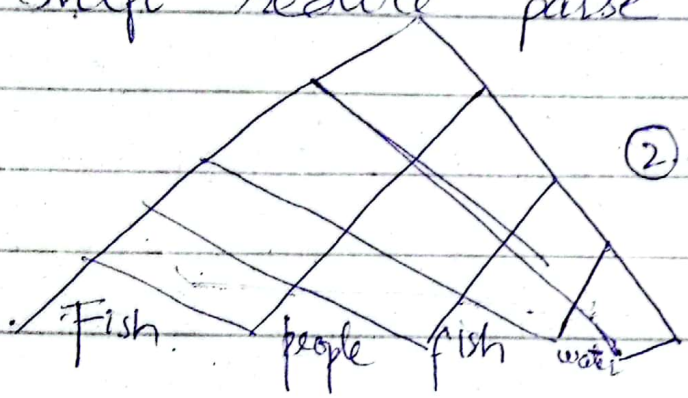
Tuesday

Chart Parsers

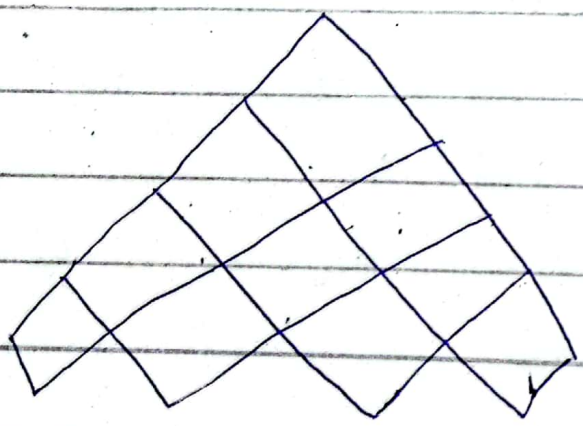
CKY parsing: (Bottom-up Parser).

→ Must have CNF for it

→ Shift reduce parse



② → right hand mein N or V ki single production.



① → right hand mein unary

VP → V → 0.1

V → people → 0

NP → N → 0.7

N → people → 0.5

NP → 0.35

V → 0.1 * 0.1

V → 0.01

NP → 0.35
V → 0.1
N → 0.5

People

VP → 0.06
NP → 0.19
V → 0.6
N → 0.2

fish

S → NP VP → exists (0.9 S → NP VP)

NP – NP

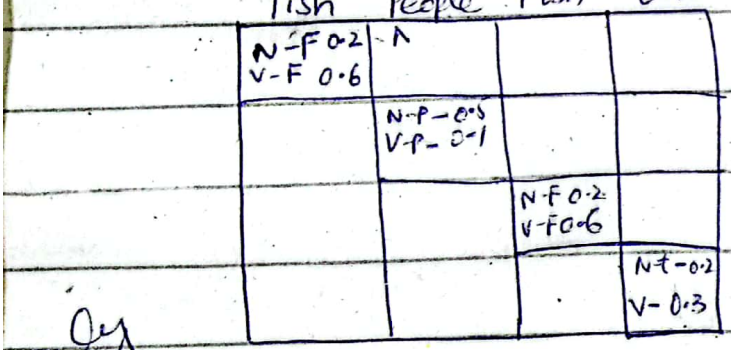
NP – V

NP – N → Not exists

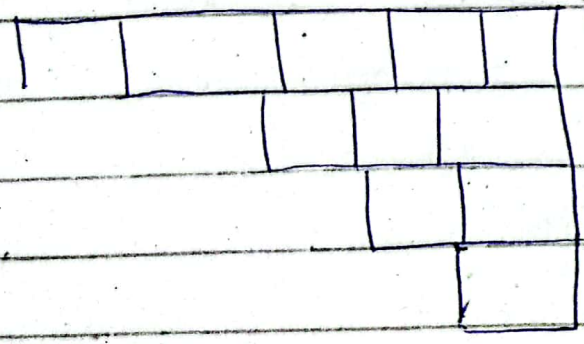
$P_1 = 0.9 * 0.35 * 0.06 \Rightarrow$

Fish People Fish Tank

fish People Fish tanks



only two combinations should exist



→ agar left do p do sy zayada exist kr rahy hun tu aik rakhen gy high prob. wali

Fish	People	fish	tank
N - 0.2 V - 0.6 NP - N - 0.14 VP - V - 0.06 S - VP - 0.006	VP - VNP - 0.5 * 0.6 * 0.35 NP - NPNP → 0.1 * 0.4 * 0.35 S - NPVP → 0.4 * 0.14 * 0.01 S - VP → 0.0105		
fish	fish people	fish people fish	people fish
	N - 0.5 V - 0.1 NP - 0.35 VP - V - 0.06 S - VP - 0.001	NP - NPNP	
	people	people fish	people fish tank
		N - 0.2 V - 0.6 NP - N - 0.14 VP - V - 0.06 S - VP - 0.006	
		fish	fish tank
			N - 0.2 V - 0.3 NP - 0.14 VP - V - 0.03 S - VP - 0.003
			tank

fish | people fish

fish people | fish

people | fish tank

people fish | tank

Evaluating Constituency Parsing

- Brackets

Gold standard → human

Candidate → Machine

Machine Translation

Embedding

Machine Translation

Embedding

V → vocabulary size of corpus

N = Embedding dimension size

W_1 = ~~VXNXV~~ NXV ⇒ weights

B / b_1 = NX1 ⇒ biasness

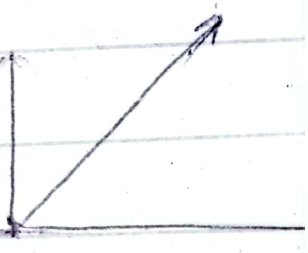
W_2 = VXN

b_2 = VX1

Softmax ⇒ values ko prob. mein conve
krta.

→ batch size must be in the form of 2's power. (2, 4, 8, 16 etc.)

* $\boxed{ReLU = \max(0, x)}$
input of hidden



Softmax → ranges values from 0 to 1 such that their sum must be equal to 1.
↓
hidden and output

Loss:

→ Loss should gradually decreas

Cross-Entropy Loss:

Friday:

22-011-24

Minimizing the Cost:

Cobow:

Output → prediction

Byproduct → Embedding

Zero Shoot.

few Shot

Pet

vs

Cat dog

Intrinsic Evaluation:

Test relationship b/w analogies

Extrinsic Evaluation:

→ external mein use kr k evaluate krna.

→ test embeddings on external task

Friday 29-11-2024

Chatbot

Factoid questions → exact answer / to the point ↗ from google

Complex (narrative) questions
→ discussions

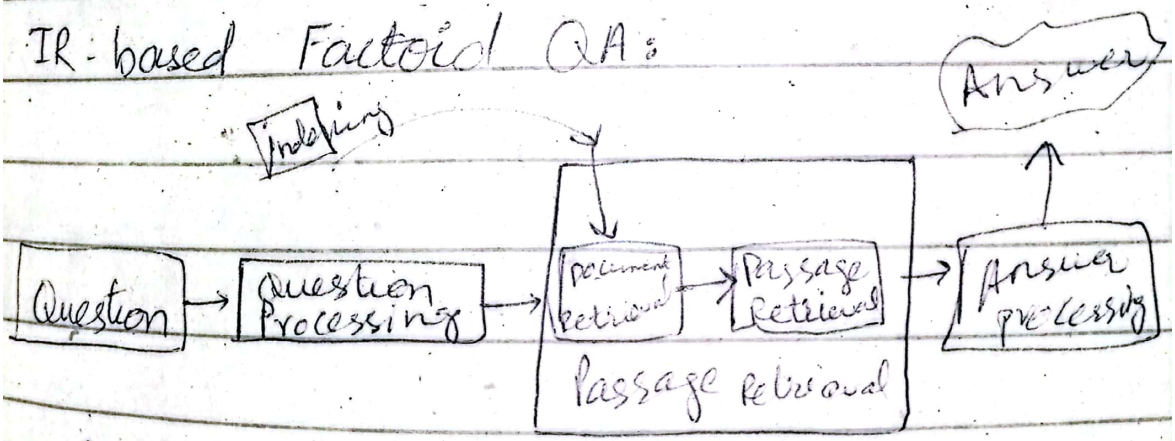
Paradigms of QA

IR-based:
→ document → Google → unstructured

Knowledge based:
→ database → Structure

Hybrid based:
→ ChatGpt, Llamah, gemini
→ Combo of both

IR-based Factoid QA:



Q.A ⇒ independent Q, A

Chatbot ⇒ depends on each other.

→ Question Processing ← Query Formulation
                      ← Answer type detection
→ Passage Retrieval
→ Answer Processing

Knowledge based:
→ Build semantic relations
(times, dates, locations, entities, numeric

→ Database can be
→ Geospatial
→ Os

Answer Type Detection:
→ decide named entity type
(person, place) of the answer

Query Formulation:
Choose query keywords for
IR System

Question Type classification:
→ definition? Math? Code?

Focus detection:
→ Find question words that are
replaced by answer

Answer type detection:
→ named entities ACL

Answer type taxonomy: (2002) Coling
→ abbreviation → entity → description
→ human → location → Numeric
* 50 finer classes

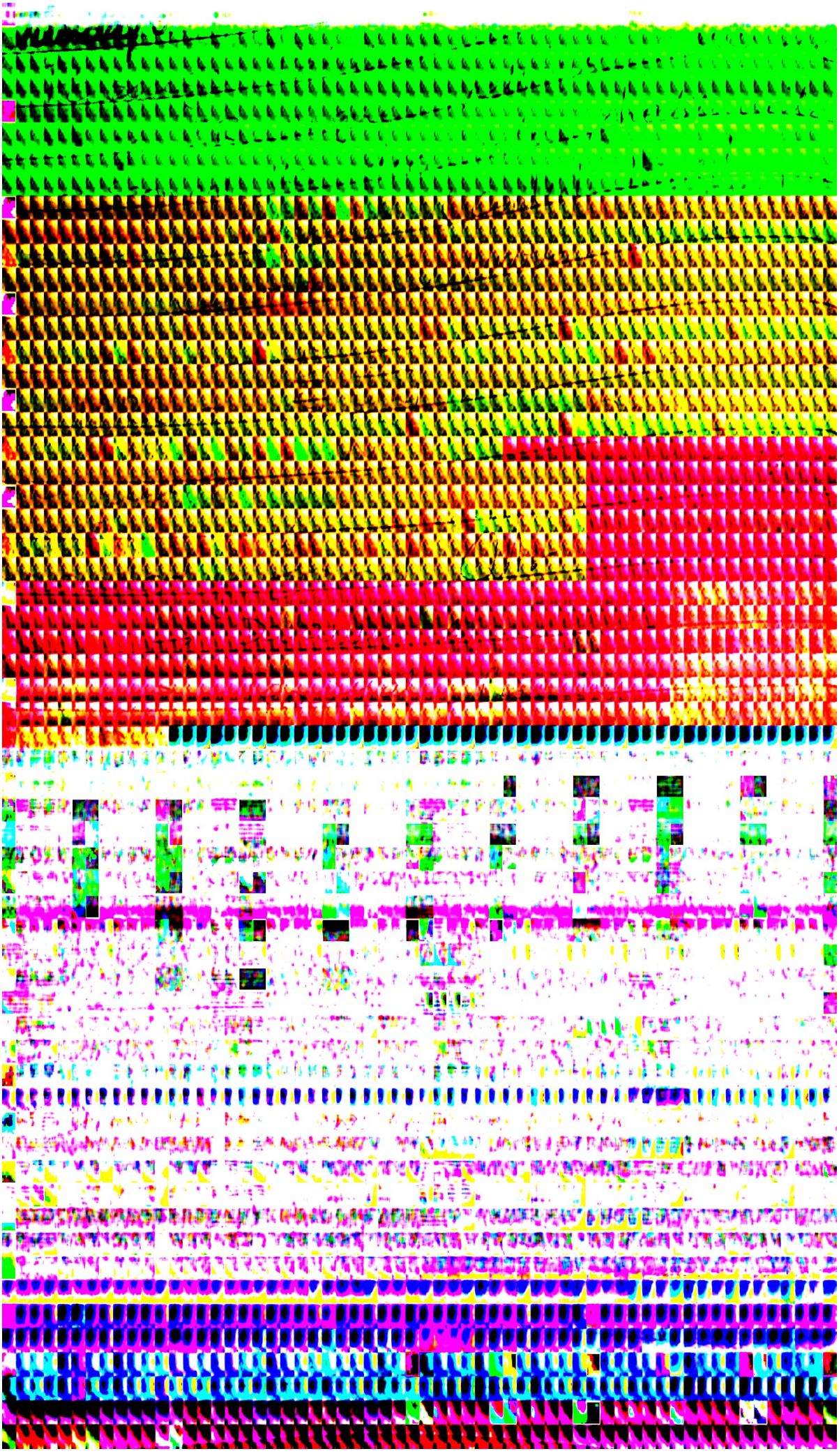
Answer Type detection:
→ Hand-Written Rules
→ ML
→ Hybrid

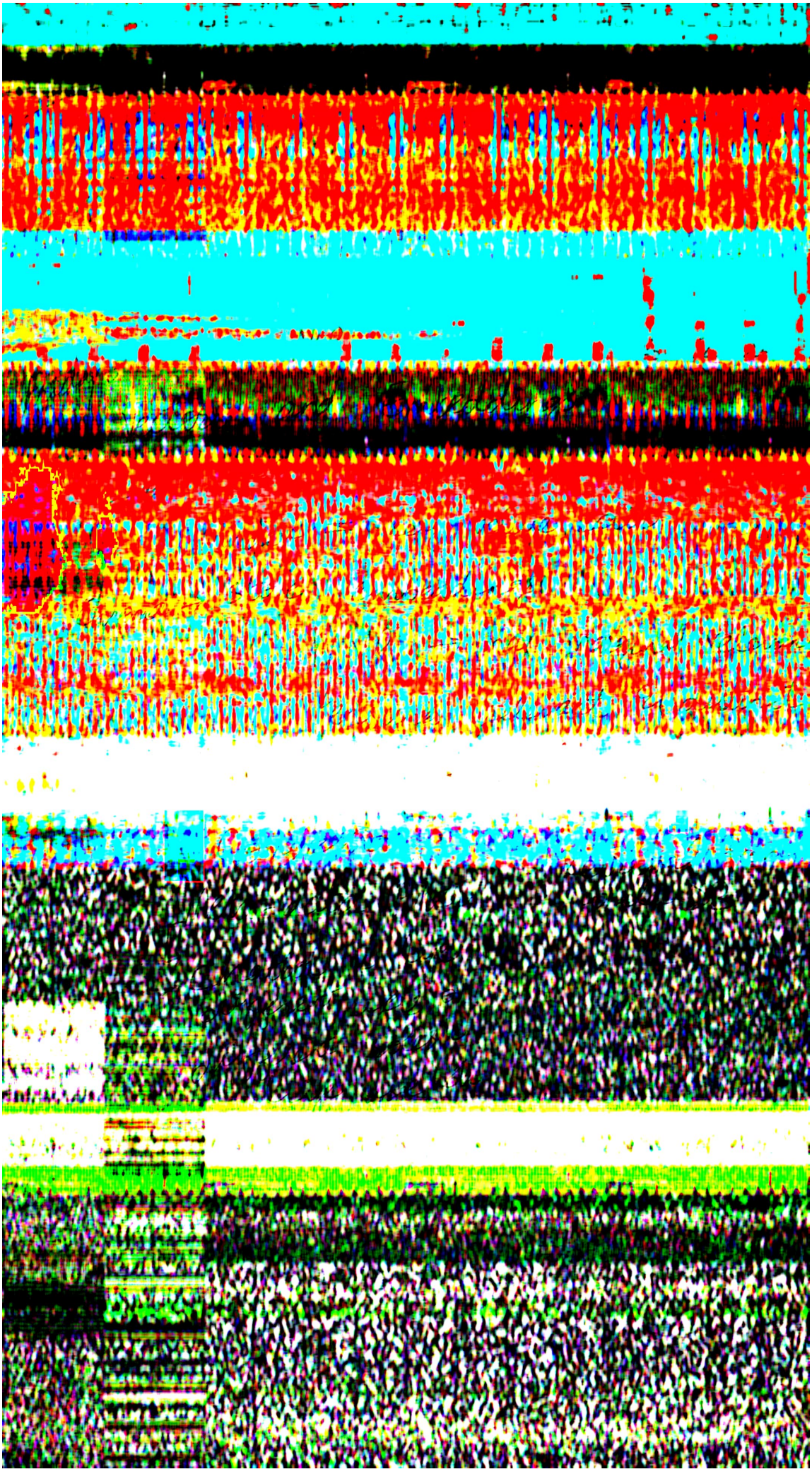
headword

Keyword Selection:
→ just remove relevant words and stopped word.

who coined the term "cyberspace" in his novel "neuromancer"?

Passage Retrieval and answer extraction
indexing ⇒ google → require data store
radix, chroma;





Neural Machine Translation

Speech to text → text to Speech

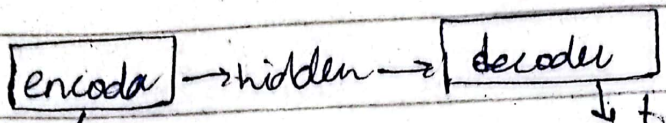
Question → Answer

ASR / STT → TTS

Seq2Seq Model:

LSTM → Long Short Term Memory

→ solves vanishing gradient problem



encoda → input or source language ko learn karta

decoder → target language ko learn karta.

English to French

Encoder → only one layer → embe

this layer has multiple LSTM

CBOW model has 3 layers

↳ input, hidden, output

→ decoder independent train hota

→ encoder dependent

learning two languages irresp

of their so? meaning?

input output size difference

its time for tea

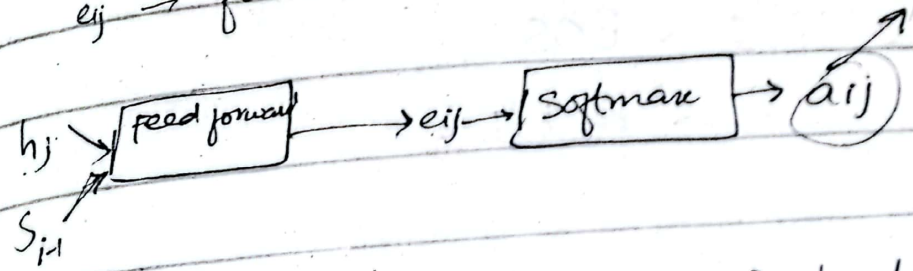
Info Bottleneck:

Fix amount of info decoder ko pass jati.

weighted k liye attention layer ko b train krain gy.

Attention Layer:

e_{ij} → function → alignment.



Decoder, Encoder → LSTM and Embeddings ke dimensions khud fix krty.

COW → ki input Content vector

Queries, key & value:

pairs mein data dein gy ← parallel Corpora → Both languages ki translation/sentence hun gy.

→ approx 1Lac sentences

Source → key

Target → value

alignment miss hoti he.

Its time for tea. ٹی ٹائم ہو گیا ہے۔

attention ki waja sy flexibility aati.

↑ layer

Machine Translation → collected

Unverified data

~~Eg.~~ Crowd sourcing

Verified → more than one person kerta

Pre-trained model

1 → EOS

0 → padding