CS626: Speech, NLP and the Web

RNN, Seq2seq, Machine Translation

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week of 9th November, 2020

Recurrent Neural Network

Acknowledgement:

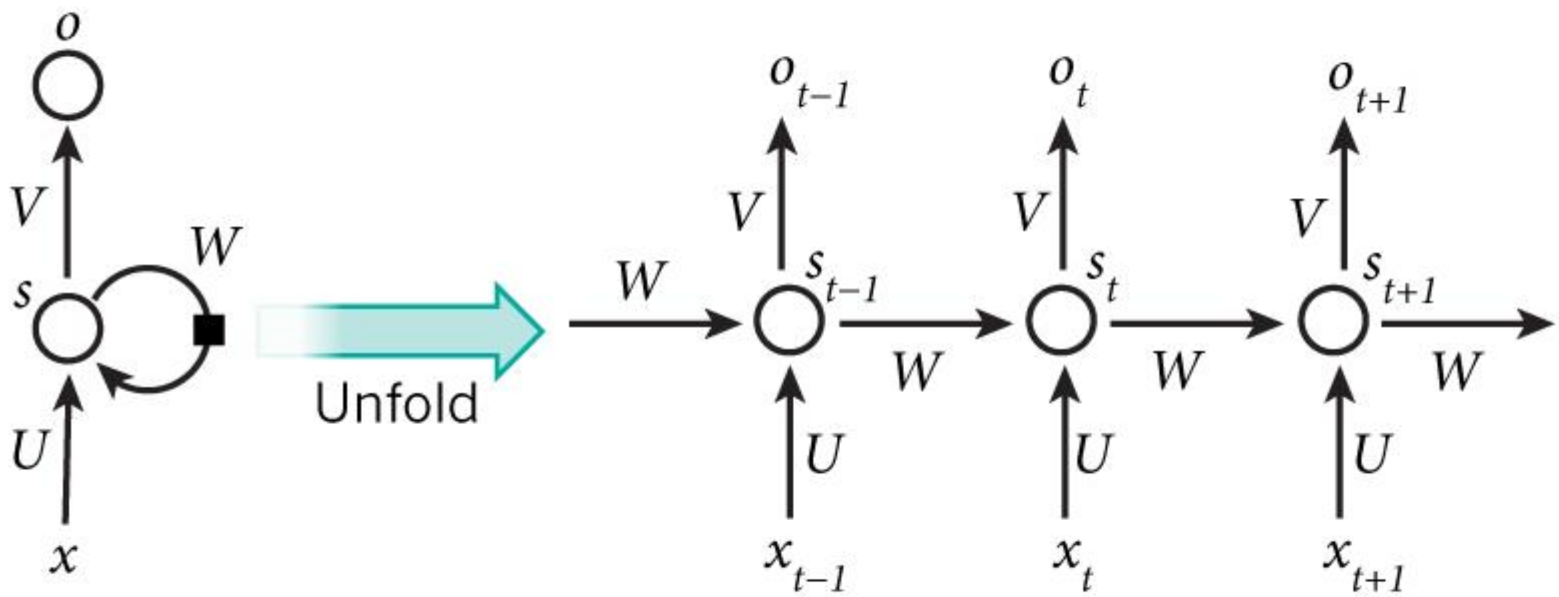
1. http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

By Denny Britz

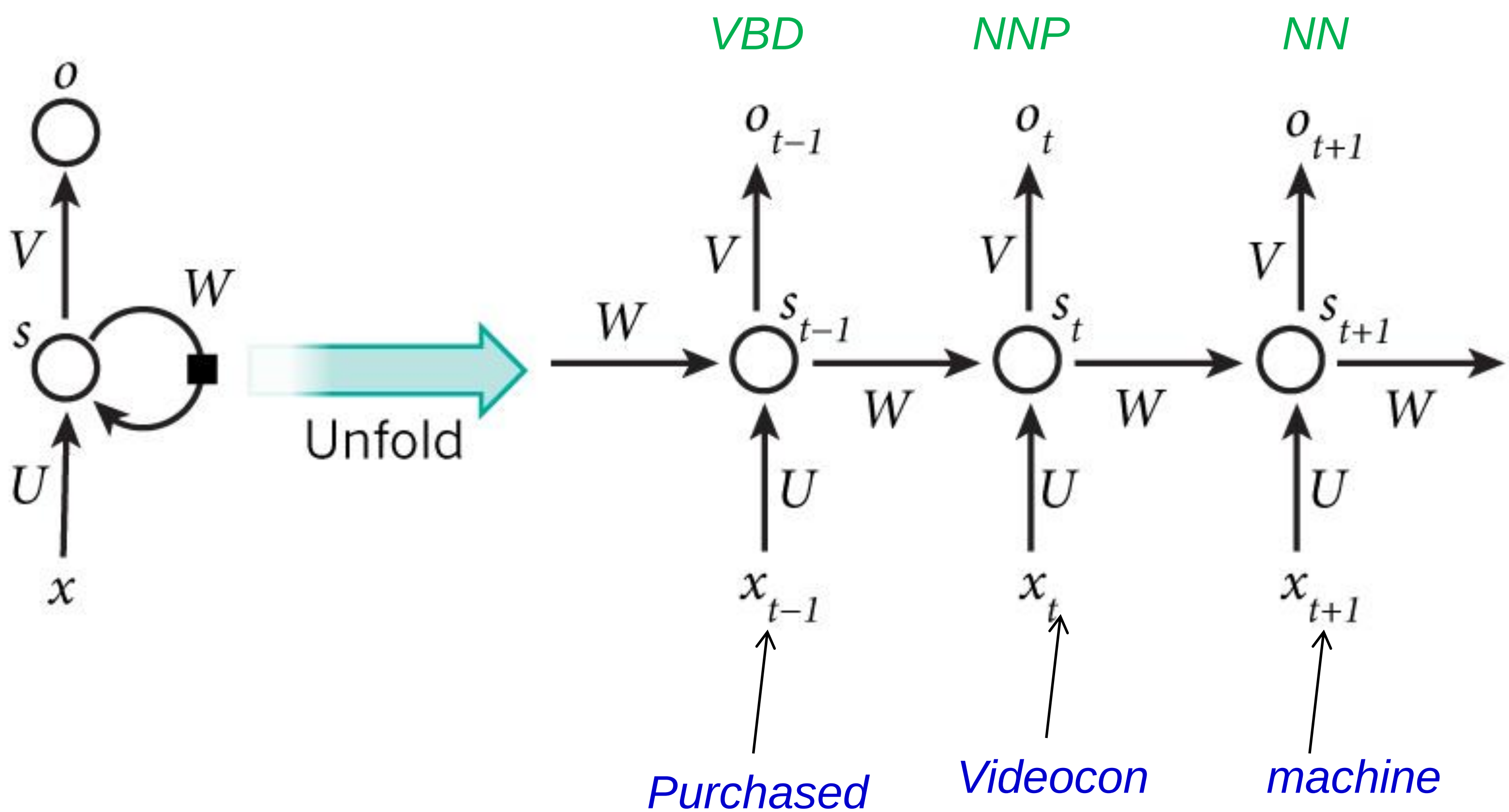
2. Introduction to RNN by Jeffrey Hinton

http://www.cs.toronto.edu/~hinton/csc2535/lectures.html

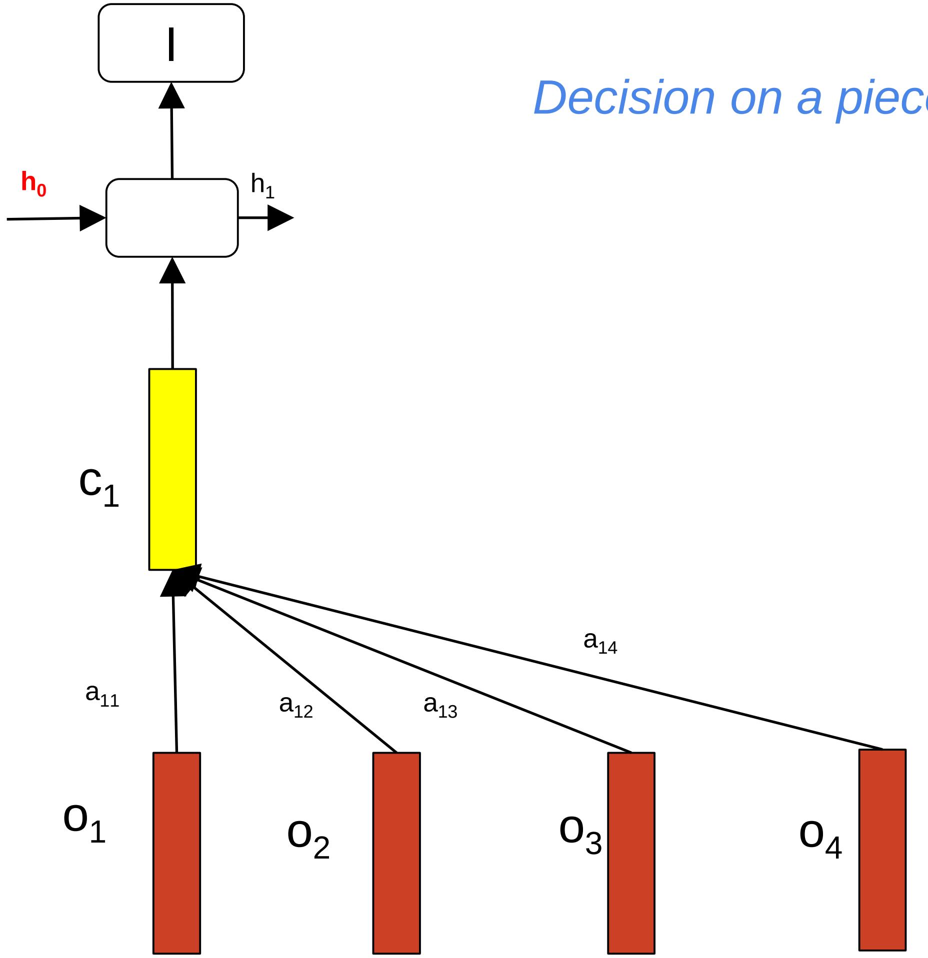
Sequence processing m/c



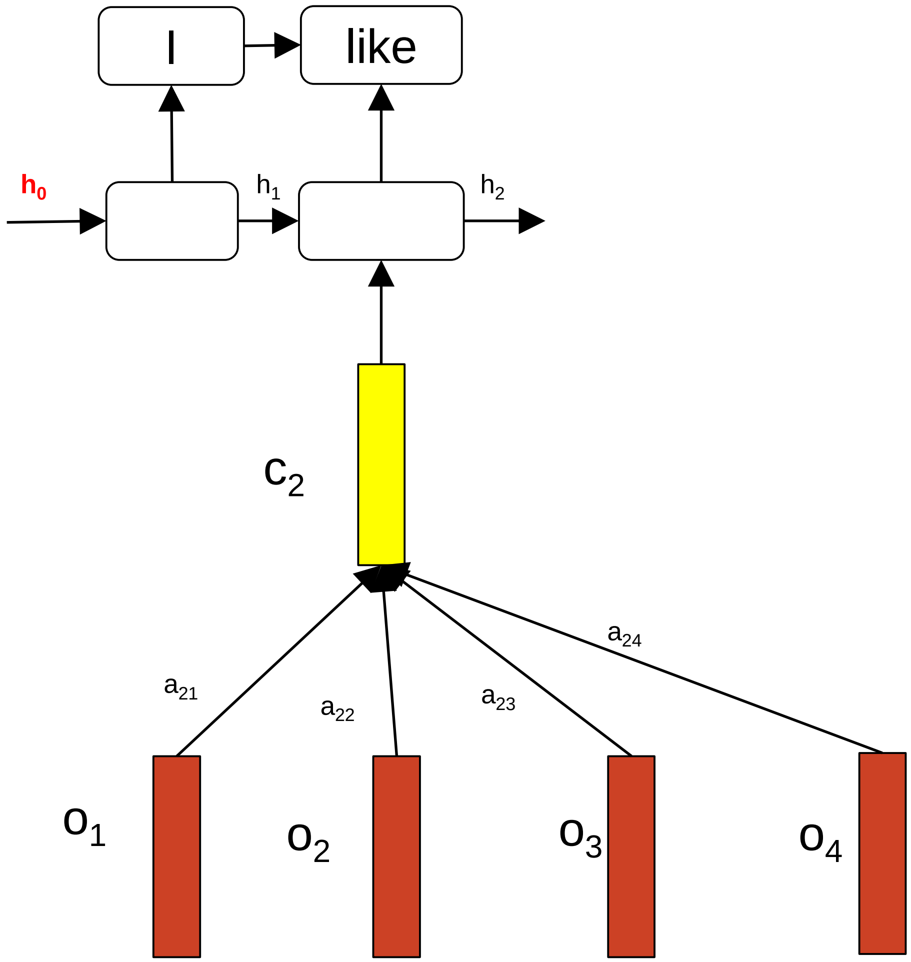
E.g. POS Tagging



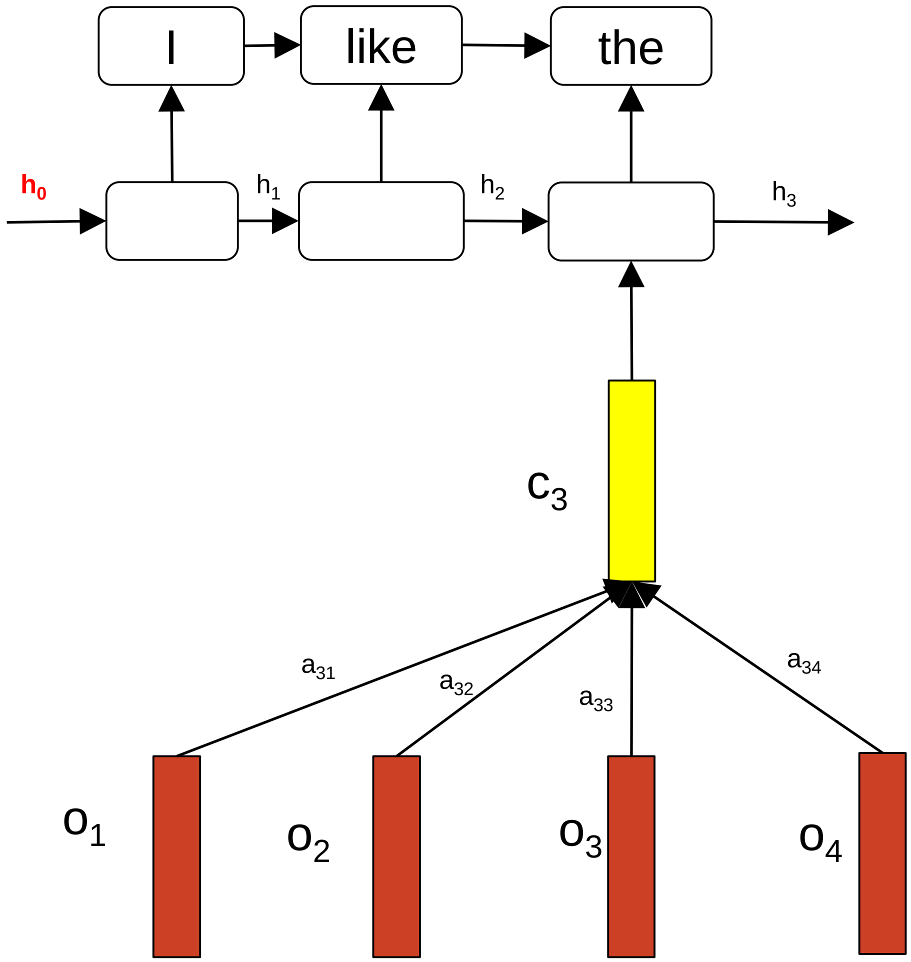
E.g. Sentiment Analysis

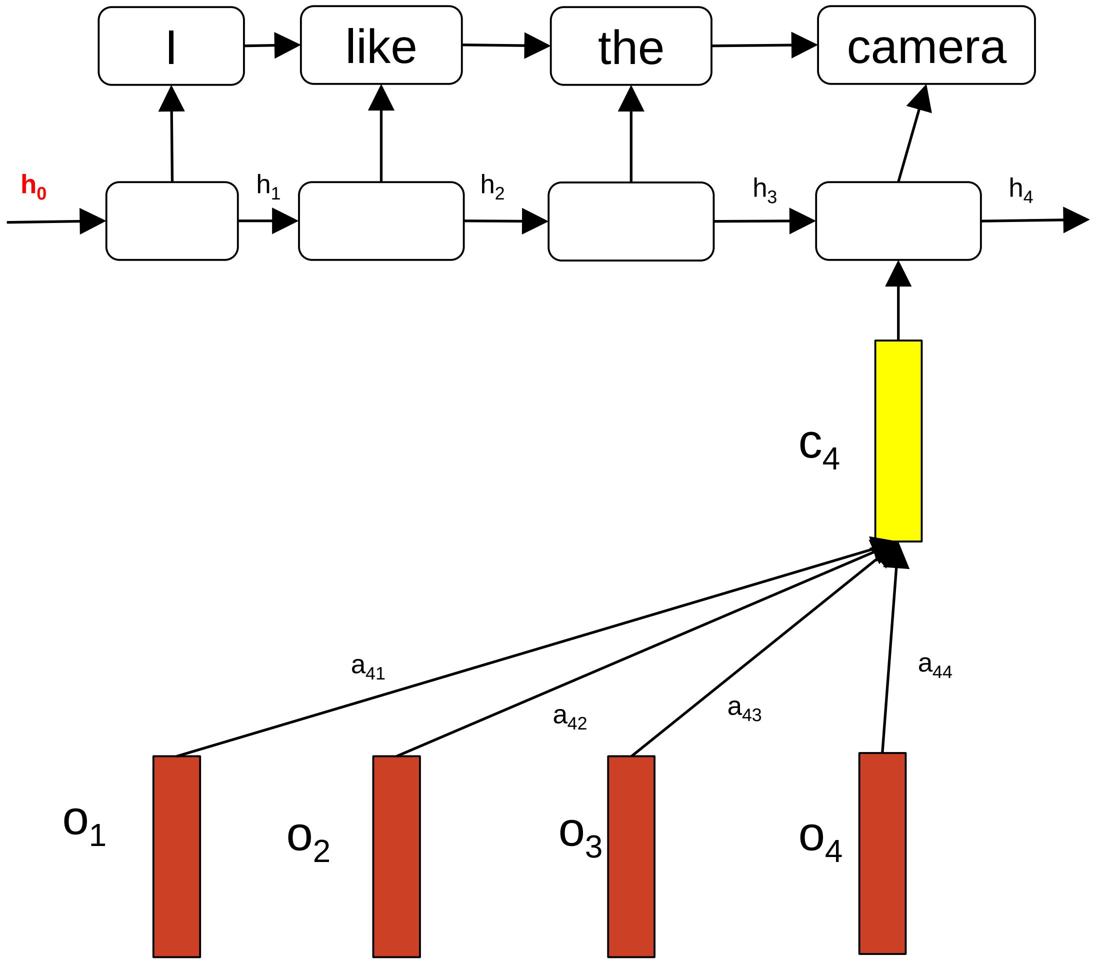


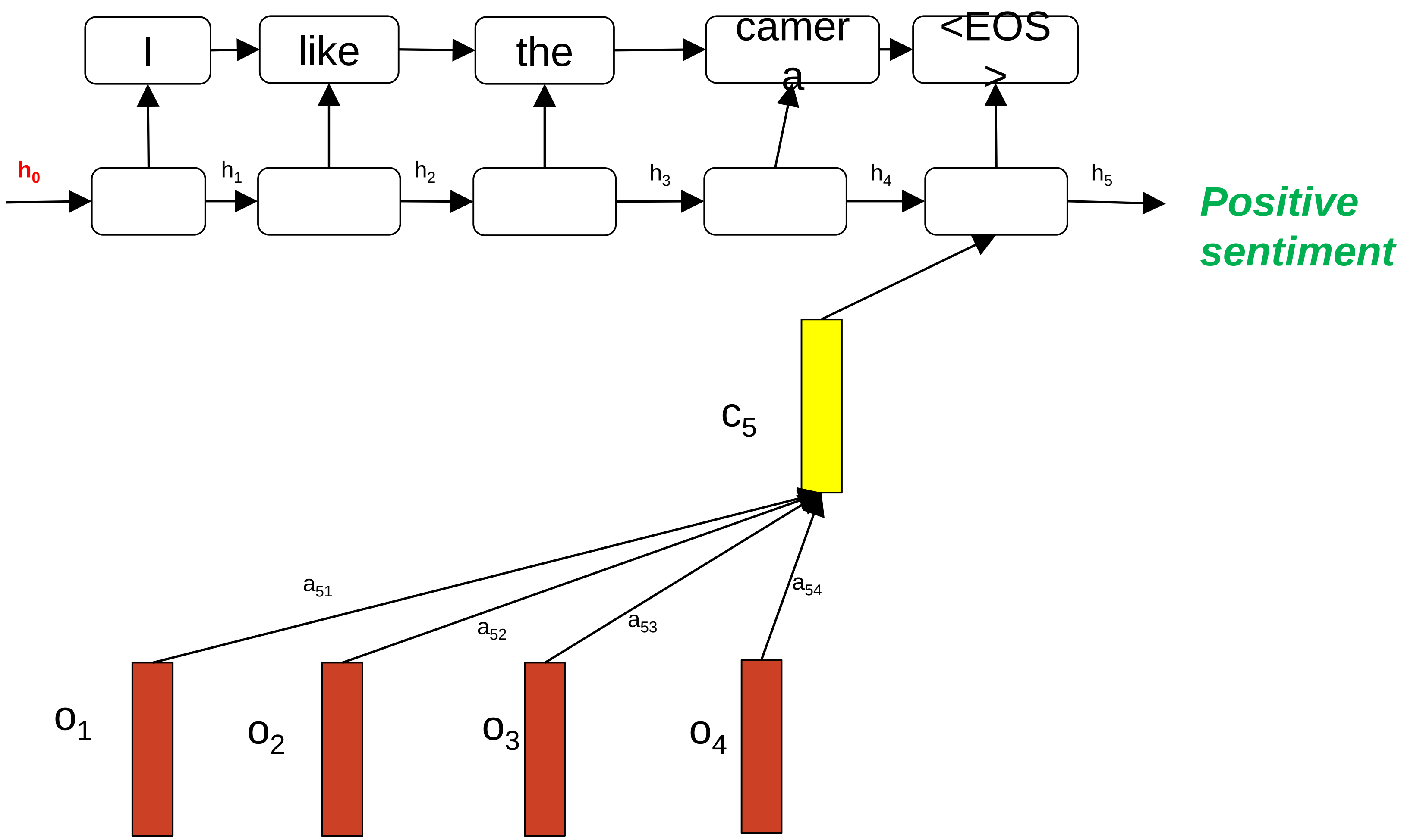
Decision on a piece of text



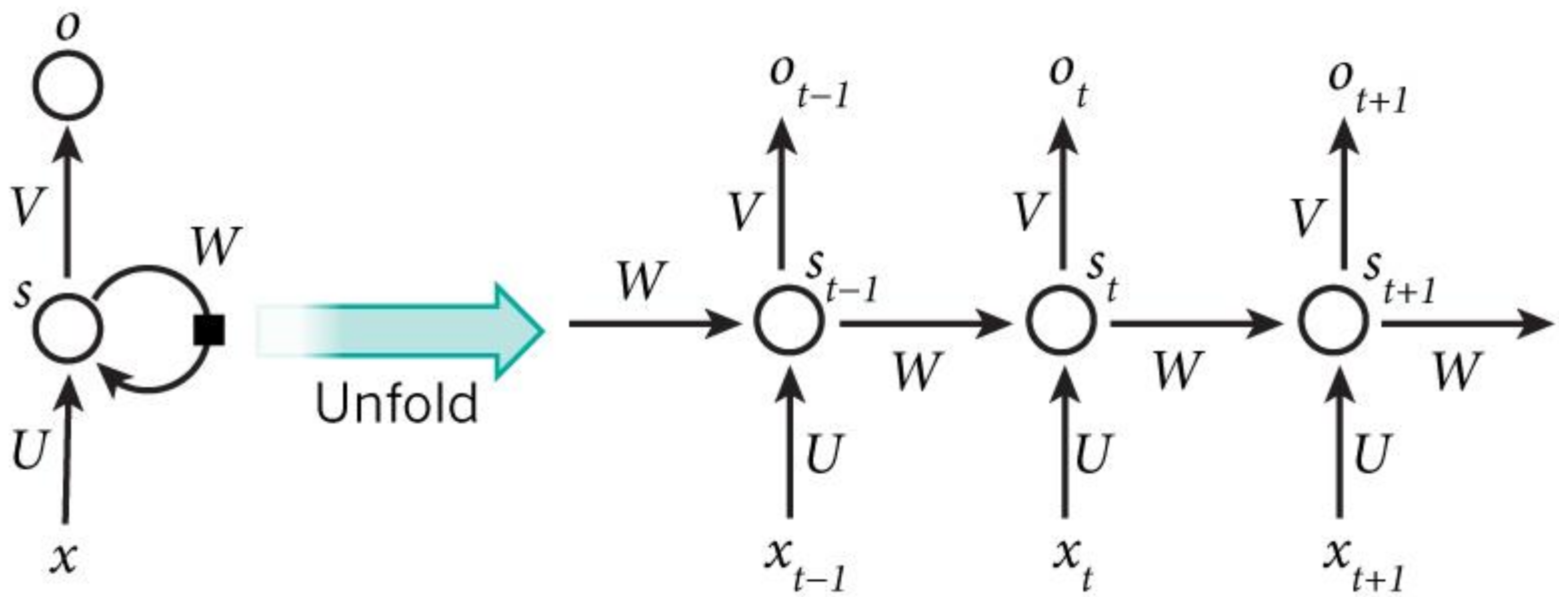
I
like
h0
h1
h2
c2
a21
a22
a23
a24
o1
o2
o3
o4


I
like
the
h_0
h_1
h_2
h_3
c_3
a_{31}
a_{32}
a_{33}
a_{34}
o_1
o_2
o_3
o_4


I
like
the
camera
h0
h1
h2
h3
h4
c4
a41
a42
a43
a44
o1
o2
o3
o4


I
like
the
camer
a
<EOS
>
h0
h1
h2
h3
h4
h5
Positive
sentiment
c5
a51
a52
a53
a54
o1
o2
o3
o4

Back to RNN model



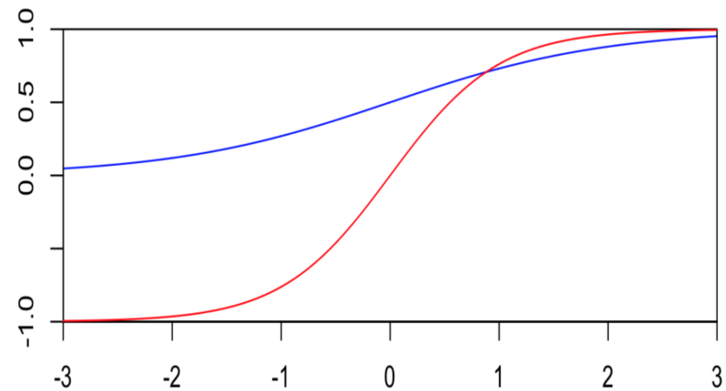
Notation: input and state

- x_t is the input at time step *t*. For example, could be a one-hot vector corresponding to the second word of a sentence.
- s_t is the hidden state at time step *t*. It is the "memory" of the network.
- $s_t = f(U.x_t+Ws_{t-1})$ ***U* and *W* matrices are learnt**
- *f* is a function of the input and the previous state
- Usually *tanh* or *ReLU* (approximated by *softplus)*

Tanh, *ReLU* (rectifier linear unit) and Softplus

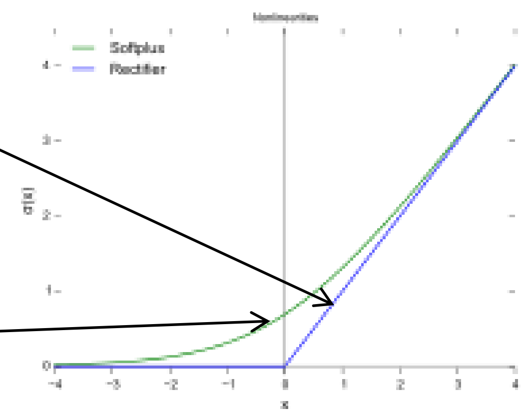
$$\tanh = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$\tanh =$



$$f(x) = \max(0, x)$$

$$g(x) = \ln(1 + e^{x})$$



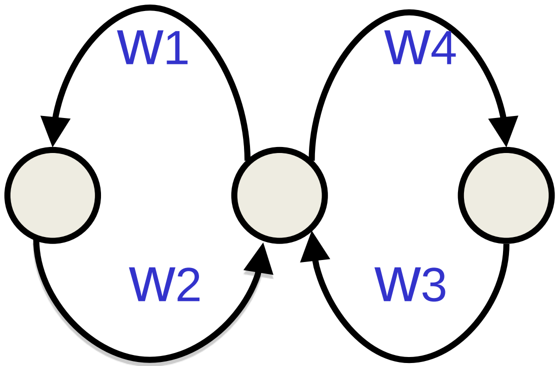
Notation: output

- o_t is the output at step t
- For example, if we wanted to predict the next word in a sentence it would be a vector of probabilities across our vocabulary
- $o_t = softmax(V.s_t)$

Operation of RNN

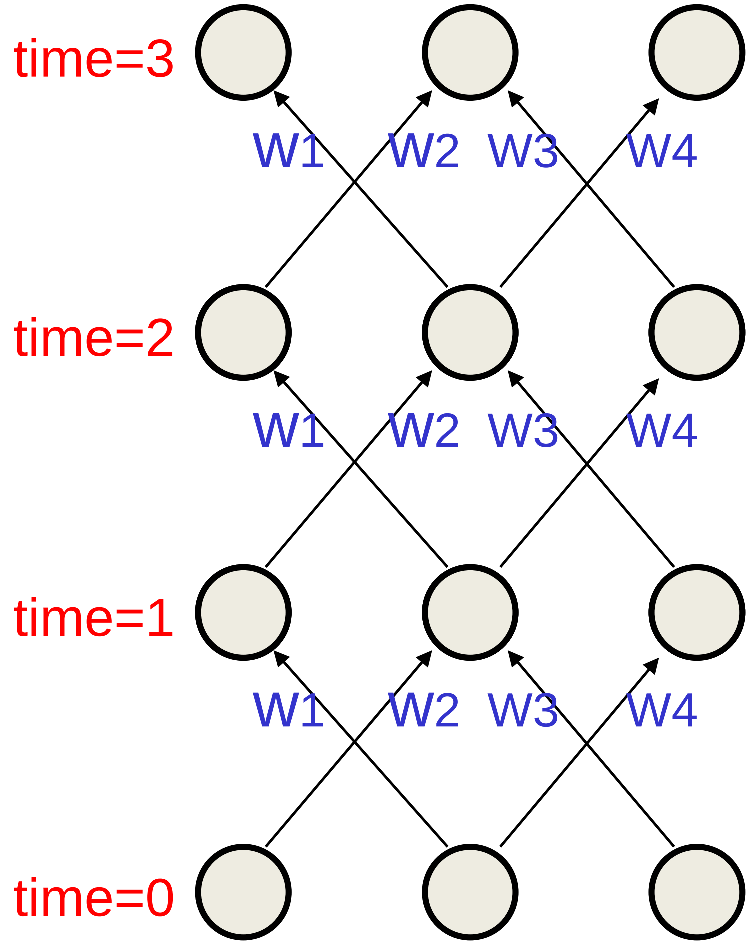
- RNN shares the same parameters (U, V, W) across all steps
- Only the input changes
- Sometimes the output at each time step is not needed: e.g., in sentiment analysis
- Main point: the hidden states !!

The equivalence between feedforward nets and recurrent nets



Assume that there is a time delay of 1 in using each connection.

The recurrent net is just a layered net that keeps reusing the same weights.



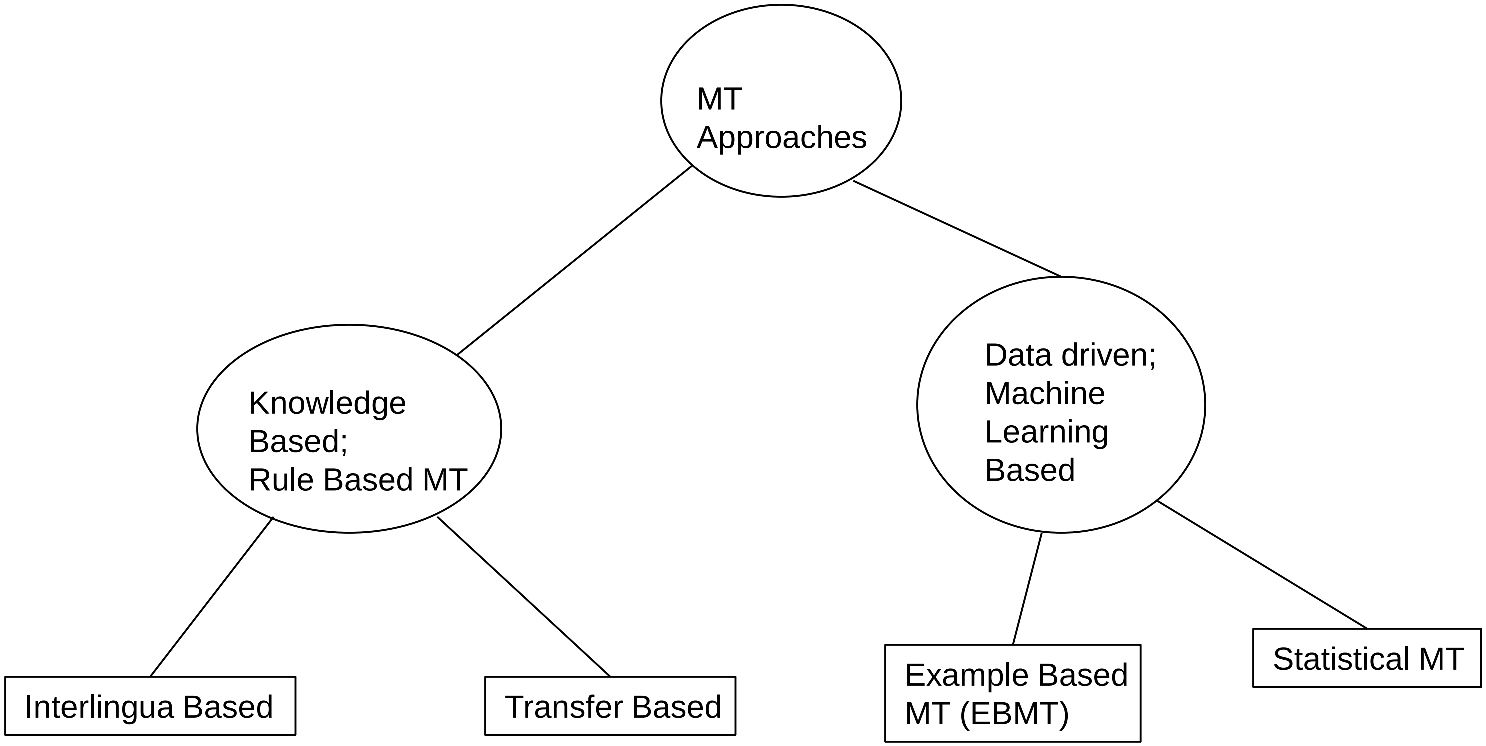
Machine Translation

(useful start: Machine Translation, Pushpak Bhattacharyya, CRC Press, 2015)

Motivation for MT

- MT: NLP Complete
- NLP: AI complete
- AI: CS complete
- How will the world be different when the language barrier disappears?
- Volume of text required to be translated currently exceeds translators’ capacity (demand > supply).
  - *Solution*: automation

Taxonomy of MT systems



Why is MT difficult?

Language divergence

Why is MT difficult: Language Divergence

- One of the main complexities of MT: *Language Divergence*
- Languages have different ways of expressing meaning
  - Lexico-Semantic Divergence
  - Structural Divergence

Our work on English-IL Language Divergence with illustrations from Hindi
(Dave, Parikh, Bhattacharyya, Journal of MT, 2002)

Languages differ in expressing thoughts: Agglutination

Finnish: “istahtaisinkohan”

English: "I wonder if I should sit down for a while“

Analysis:

- ist + "sit", verb stem
- ahta + verb derivation morpheme, "to do something for a while"
- isi + conditional affix
- n + 1st person singular suffix
- ko + question particle
- han a particle for things like reminder (with declaratives) or "softening" (with questions and imperatives)

Language Divergence Theory:

Lexico-Semantic Divergences (few examples)

- Conflational divergence
  - F: vomir; E: to be sick
  - E: *stab;* H: *chure se maaranaa (knife-with hit)*
  - S: *Utrymningsplan;* E: *escape plan*

- Categorial divergence
  - Change is in POS category:

  - *The play is on_PREP (vs. The play is Sunday)*
  - *Khel chal_rahaa_haai_VM (vs. khel ravivaar ko haai)*

Language Divergence Theory:

Structural Divergences

- SVO→SOV
  - E: *Peter plays basketball*
  - H: *piitar basketball kheltaa haai*

- Head swapping divergence
  - E: *Prime Minister of India*
  - H: *bhaarat ke pradhaan mantrii (India-of Prime Minister)*

Language Divergence Theory: *Syntactic Divergences* (few examples)

- Constituent Order divergence
  - E: *Singh, the PM of India, will address the nation today*
  - H: *bhaarat ke pradhaan mantrii, singh, … (India-of PM, Singh…)*
- Adjunction Divergence
  - E: *She will visit here in the summer*
  - H: *vah yahaa garmii meM aayegii (she here summer-in will come)*
- Preposition-Stranding divergence
  - E: *Who do you want to go with?*
  - H: *kisake saath aap jaanaa chaahate ho? (who with…)*

Latency concerns: What is Latency?

- Example
    - Purchased videocon machine. (VBD NNP NN) (VP)
    - वीडियोकॉन मशीन खरीदी।
    - Videocon machine kharidi
- **Latency**
    - Purchased videocon machine: Verb phrase
    - English: Head initial (**Purchased** in the beginning of the phrase)
    - Hindi: Head final (**kharidi** in the end of the phrase)
    - In speech to speech translation or interactive machine translation
        - Translation of **purchased** can not be produced immediately after seeing the input string, it needs to be hold back (This phenomenon is known as **latency**)

Monotonicity

- Isolate phrases in the sentence whose translation have to be done together
- Move from one group of words to another without going back, without any regression.
- How translators translate?
  - Approach1
    - Make groups
      - Groups: I saw immediately the blue sky
    - These groups (chunks) are translated and reordered to make the final translation.
  - Approach2
    - Rearrange the sentence first keeping the target language in mind, then translate.
    - I the blue sky saw immediately.
    - Maine neela asman ko turant dekha.

Exercise

Phrase movement versus local translation,
which one should be done earlier?

Vauquois Triangle

Kinds of MT Systems

(point of entry from source to the target text)

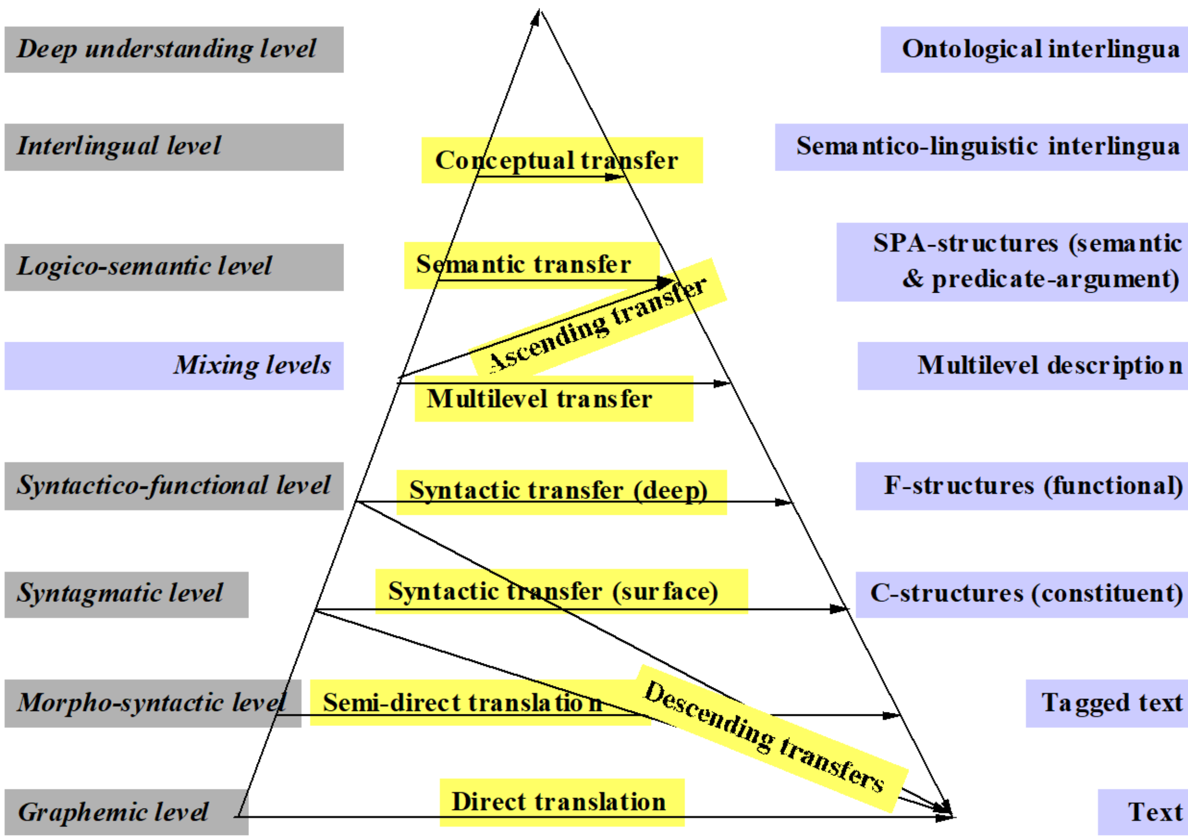
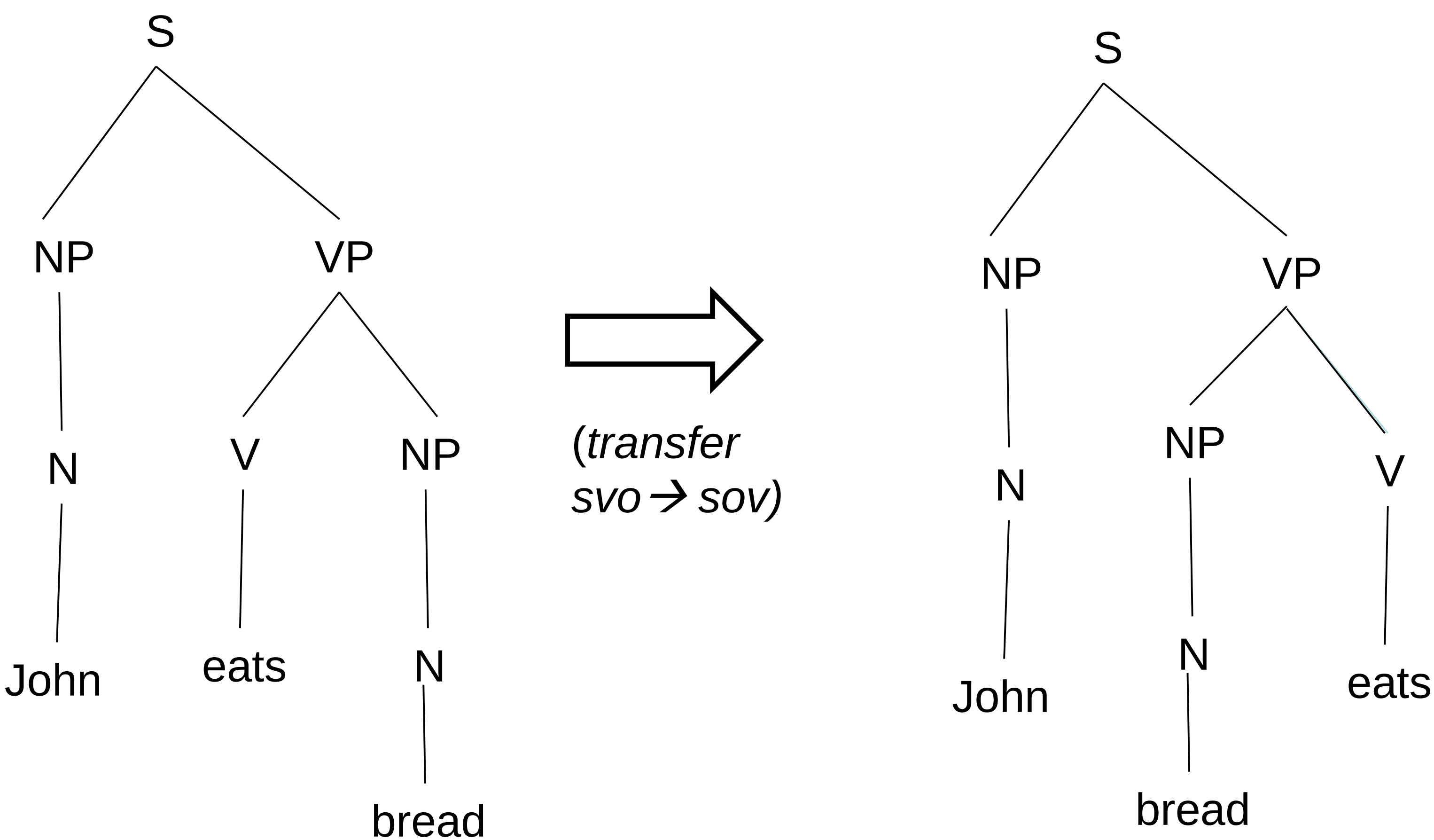


Illustration of transfer SVO→SOV



Fundamental processes in Machine Translation

- **Analysis**
  - Analysis of the source language to represent the source language in more disambiguated form
    - Morphological segmentation, POS tagging, chunking, parsing, discourse resolution, pragmatics etc.
- **Transfer**
  - Knowledge transfer from one language to another
  - Example: SOV to SVO conversion
- **Generation**
  - Generate the final target sentence
  - Final output is text, intermediate representations can include F-structures, C-structures, tagged text etc.

Universality hypothesis

Universality hypothesis: At the level of "deep meaning", all texts are the "same", whatever the language.

Understanding the Analysis-Transfer-Generation over Vauquois triangle (1/4)

H1.1: सरकार_ने चुनावो_के_बाद मुंबई में करों_के_माध्यम_से अपने राजस्व_को बढ़ाया |

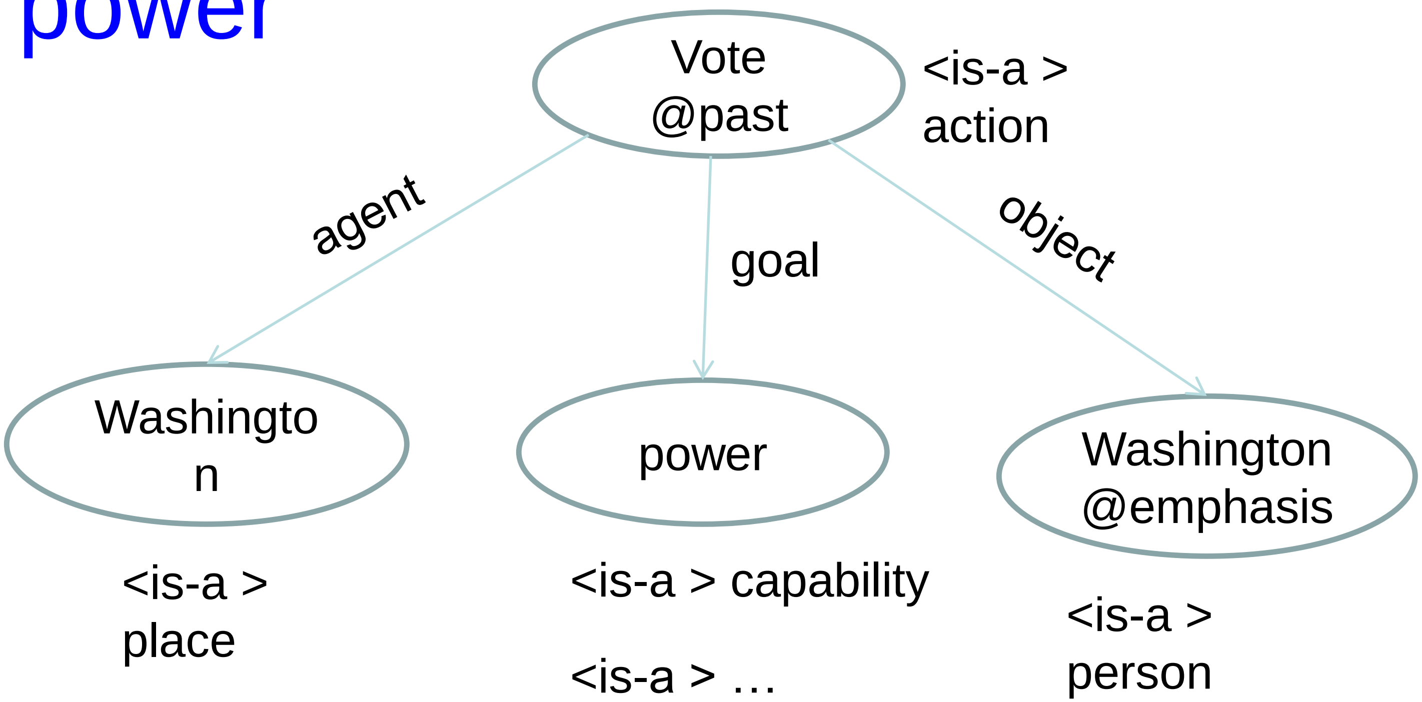
T1.1: Sarkaar ne chunaawo ke baad Mumbai me karoM ke maadhyam se apne raajaswa ko badhaayaa

G1.1: Government_(ergative) elections_after Mumbai_in taxes_through its revenue_(accusative) increased

E1.1: The Government increased its revenue after the elections through taxes in Mumbai

Interlingual representation: complete disambiguation

- Washington voted Washington to power



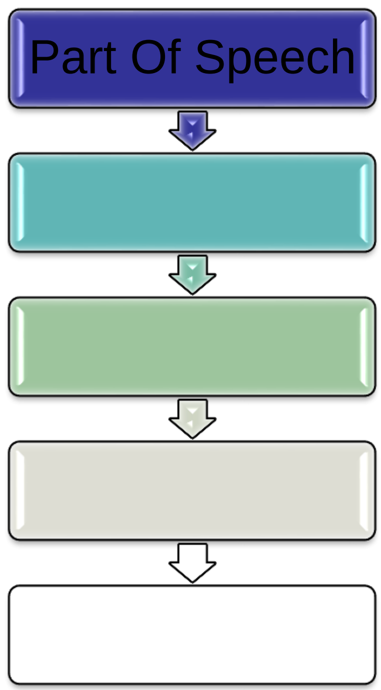
Kinds of disambiguation needed for a complete and correct interlingua graph

- N: Name
- P: POS
- A: Attachment
- S: Sense
- C: Co-reference
- R: Semantic Role

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES



Noun or Verb

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES

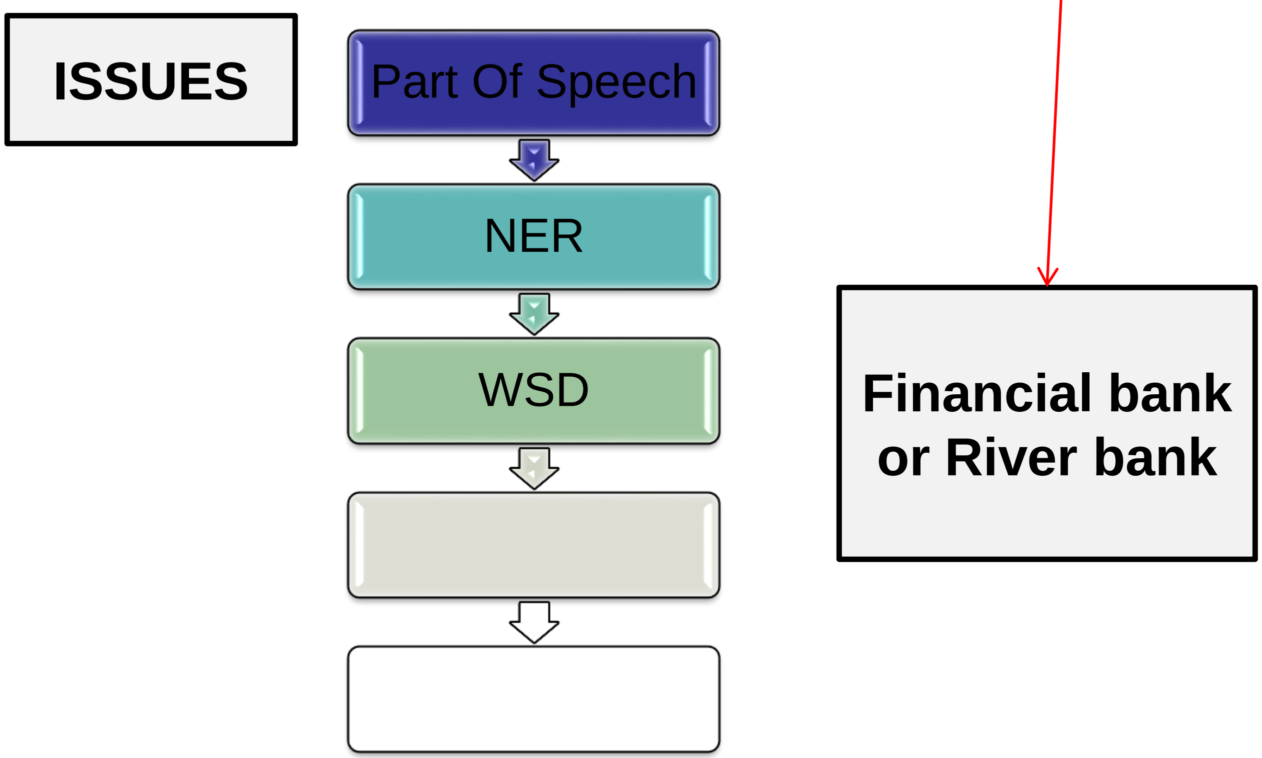
Part Of Speech

NER

John is the name of a PERSON

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*



Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES

Part Of Speech

NER

WSD

Co-reference

"it" → "bank" .

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES

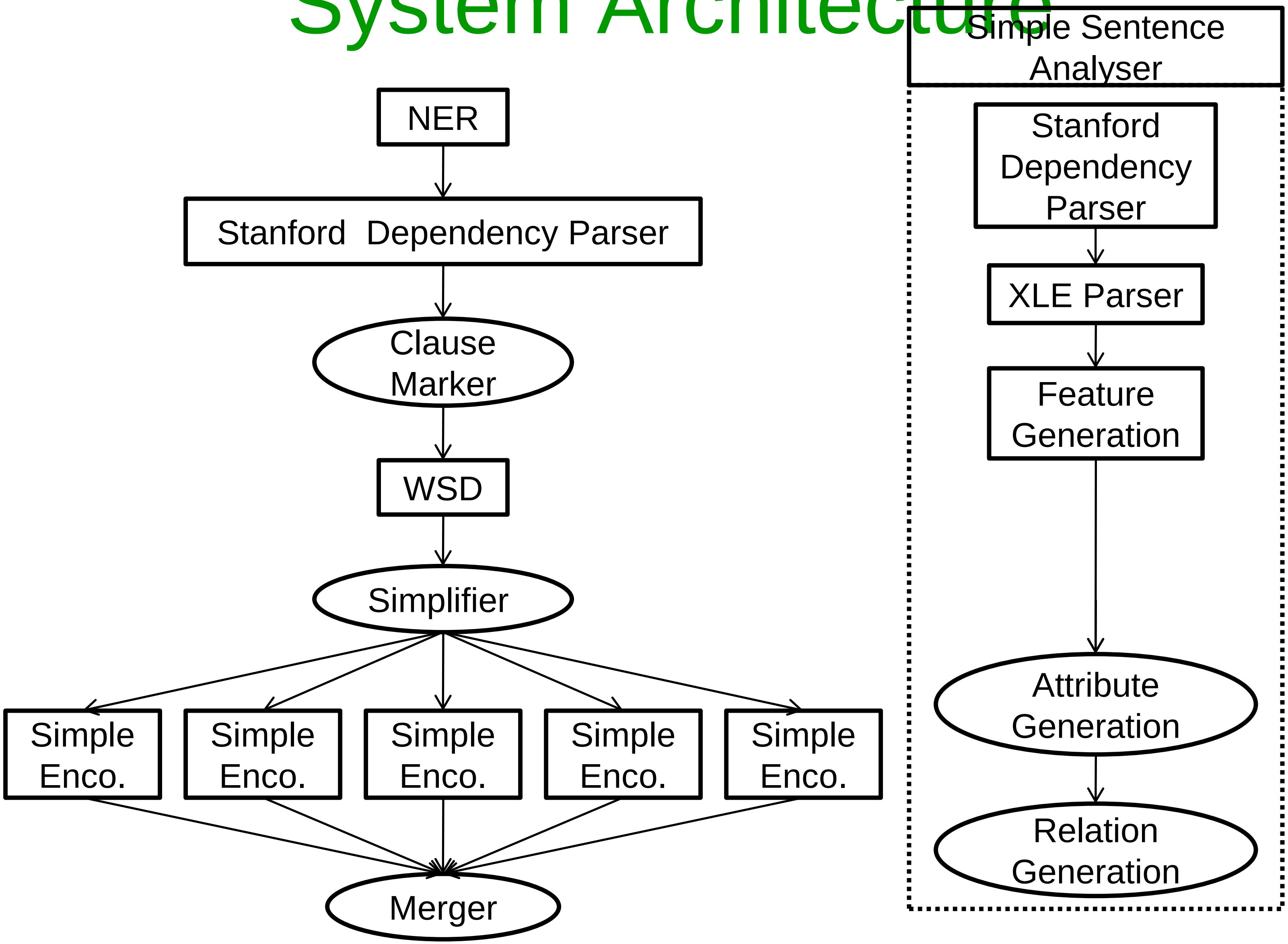
- Part Of Speech
- NER
- WSD
- Co-reference
- Subject Drop

Pro drop
(subject "I")

Typical NLP tools used

- POS tagger
- Stanford Named Entity Recognizer
- Stanford Dependency Parser
- XLE Dependency Parser
- Lexical Resource
  - WordNet
  - Universal Word Dictionary (UW++)

System Architecture



Target Sentence Generation from interlingua

Target Sentence Generation

Lexical Transfer

(Word/Phrase Translation)

Morphological Synthesis

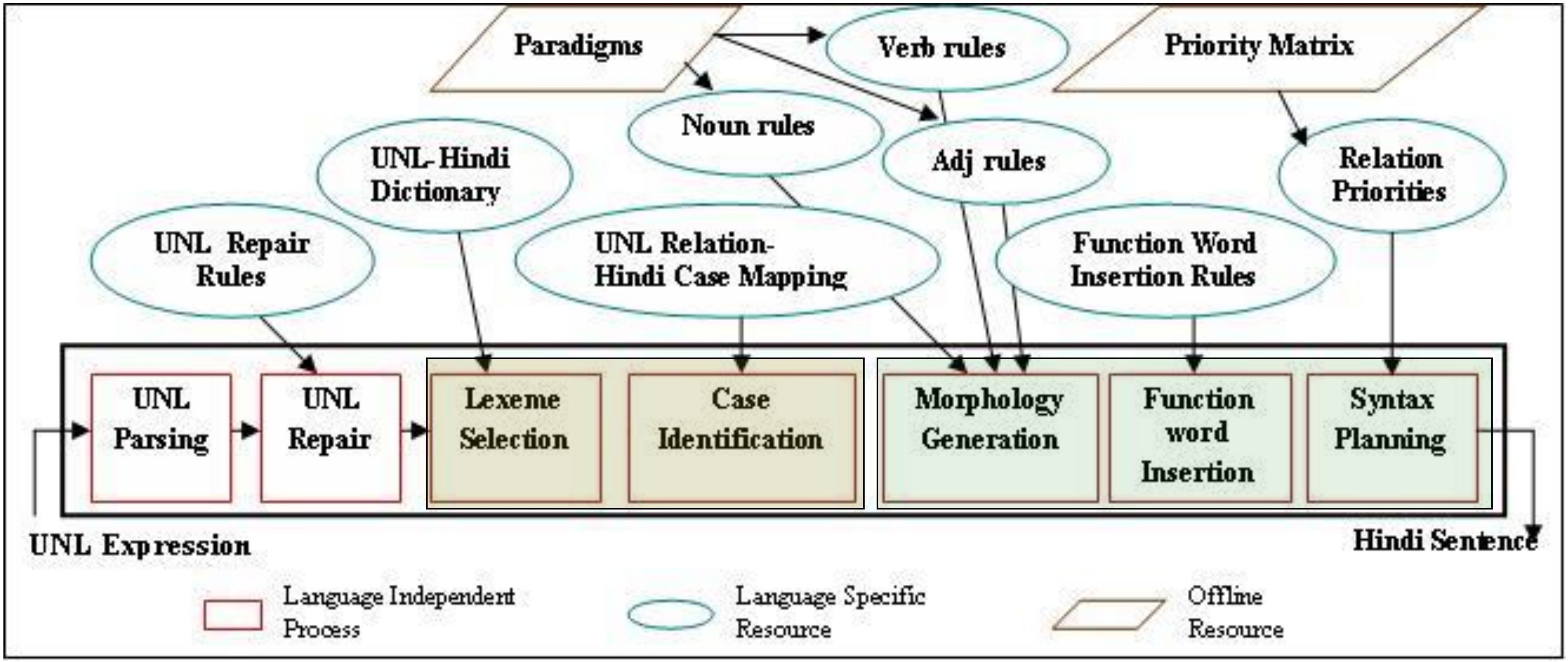
(Word form Generation)

Syntax Planning

(Sequence)

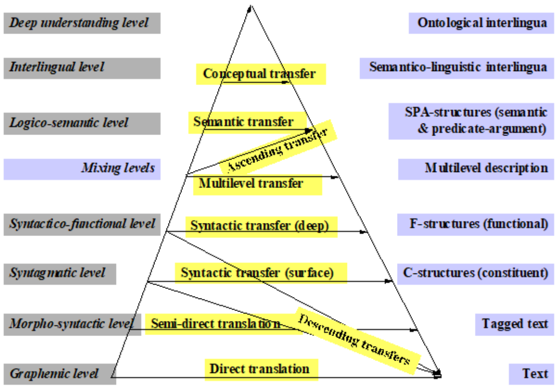
Generation Architecture

Deconversion = Transfer + Generation



Transfer Based MT

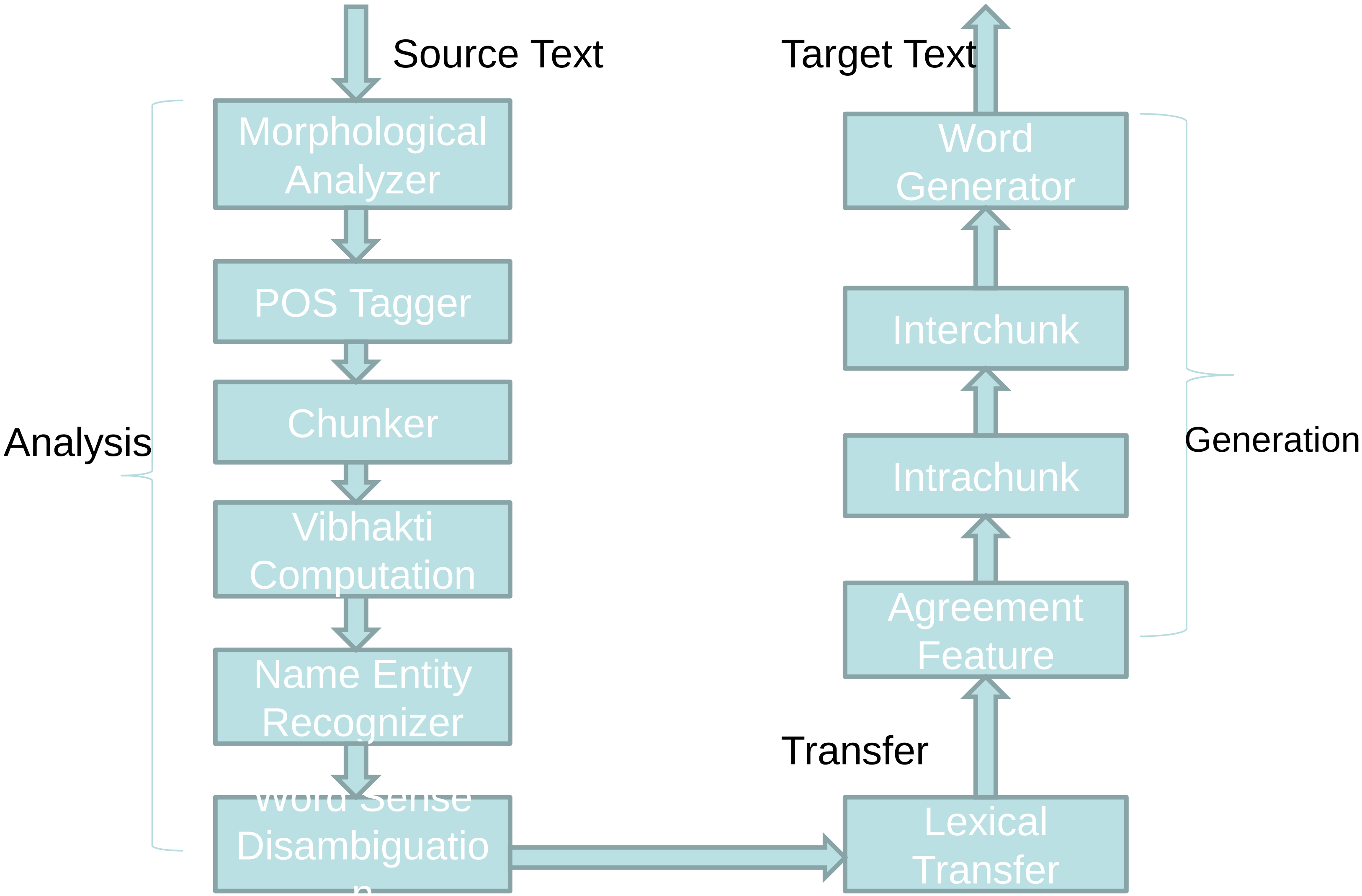
Marathi-Hindi



Indian Language to Indian Language Machine Translation (ILILMT)

- Bidirectional Machine Translation System
- Developed for nine Indian language pairs
- Approach:
  - Transfer based
  - Modules developed using both rule based and statistical approach

Architecture of ILILMT System



M-H MT system: Evaluation

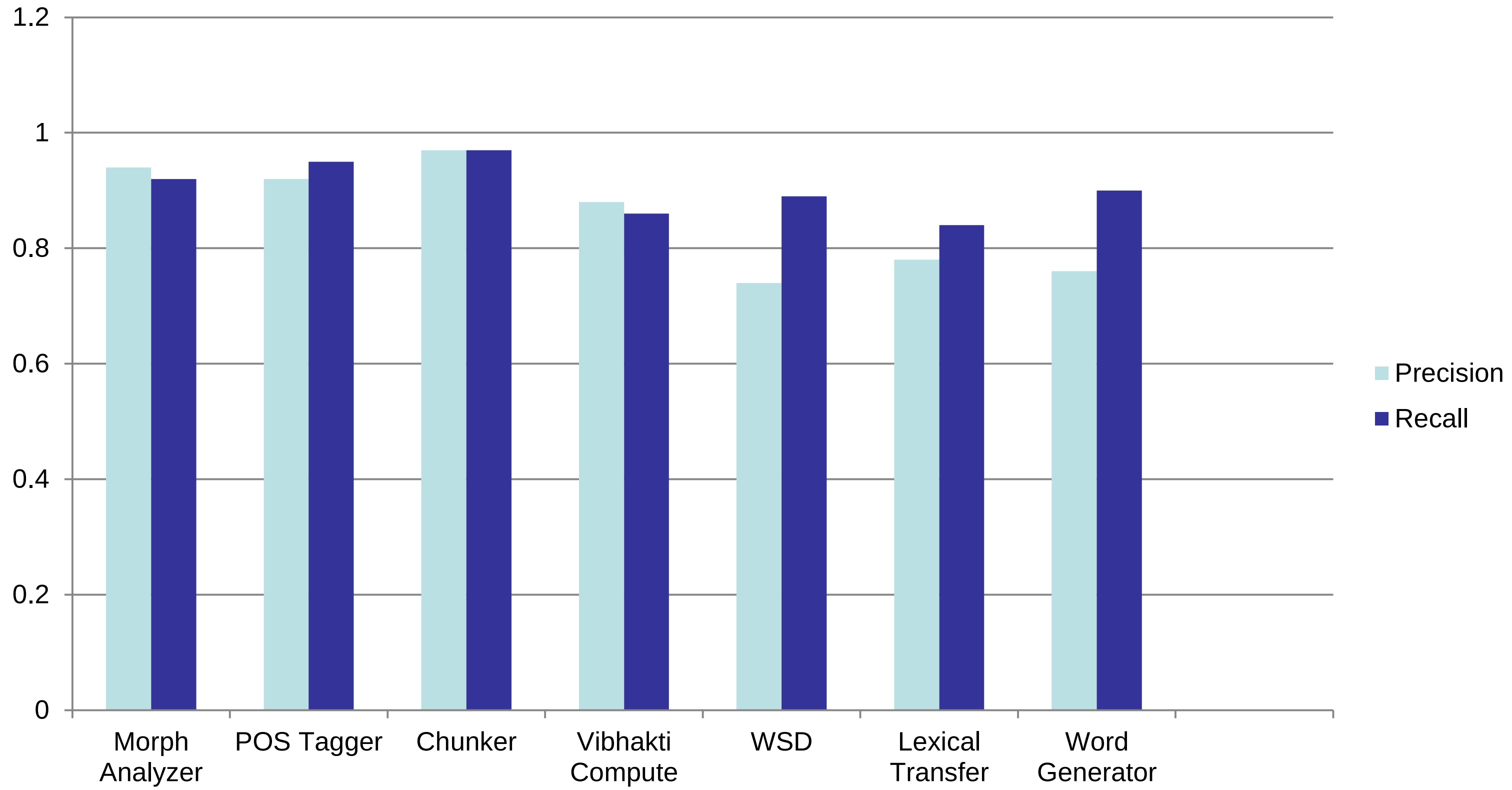
- Subjective evaluation based on machine translation quality
- Accuracy calculated based on score given by linguists

Score : 5	Correct Translation
Score : 4	Understandable with minor errors
Score : 3	Understandable with major errors
Score : 2	Not Understandable
Score : 1	Non sense translation

S5: Number of score 5 Sentences,
S4: Number of score 4 sentences,
S3: Number of score 3 sentences,
N: Total Number of sentences

$$\text{Accuracy} = \frac{1 * S5 + 0.8 * S4 + 0.6 * S3}{N}$$

Evaluation of Marathi to Hindi MT System



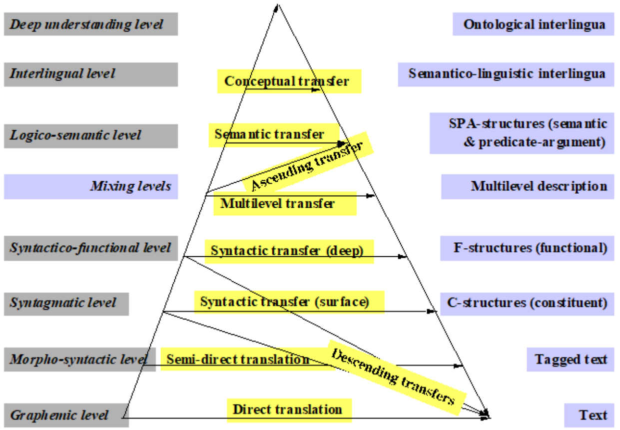
Module-wise precision and recall

Evaluation of Marathi to Hindi MT System (cont..)

- Subjective evaluation on translation quality
  - Evaluated on 500 web sentences
  - Accuracy calculated based on score given according to the translation quality.
  - Accuracy: **65.32 %**

- Result analysis:
  - Morph, POS tagger, chunker gives more than 90% precision but Transfer, WSD, generator modules are below 80% hence degrades MT quality.
  - Also, morph disambiguation, parsing, transfer grammar and FW disambiguation modules are required to improve accuracy.

Statistical Machine Translation



Czeck-English data

• [nesu]	“I carry”
• [ponese]	“He will carry”
• [nese]	“He carries”
• [nesou]	“They carry”
• [yedu]	“I drive”
• [plavou]	“They swim”

To translate …

- I will carry.
- They drive.
- He swims.
- They will drive.

Hindi-English data

• [DhotA huM]	“I carry”
• [DhoegA]	“He will carry”
• [DhotA hAi]	“He carries”
• [Dhote hAi]	“They carry”
• [chalAtA huM]	“I drive”
• [tErte hEM]	“They swim”

Bangla-English data

• [bai]	“I carry”
• [baibe]	“He will carry”
• [bay]	“He carries”
• [bay]	“They carry”
• [chAlAi]	“I drive”
• [sAMtrAy]	“They swim”

To translate … (repeated)

- I will carry.
- They drive.
- He swims.
- They will drive.

Foundation

- Data driven approach
- Goal is to find out the English sentence *e* given foreign language sentence *f* whose *p(e|f)* is maximum.

$$\tilde{e} = \underset{e \in e^*}{\operatorname{argmax}}\, p(e|f) = \underset{e \in e^*}{\operatorname{argmax}}\, p(f|e)p(e)$$

- Translations are generated on the basis of statistical model
- Parameters are estimated using bilingual parallel corpora

SMT: Language Model

- To detect *good* English sentences
- Probability of an English sentence $w_1 w_2 \ldots\ldots w_n$ can be written as

$$Pr(w_1 w_2 \ldots\ldots w_n) = Pr(w_1) * Pr(w_2|w_1) * \ldots * Pr(w_n|w_1 w_2 \ldots w_{n-1})$$

- Here $Pr(w_n|w_1 w_2 \ldots w_{n-1})$ is the probability that word w_n follows word string $w_1 w_2 \ldots w_{n-1}$.
  - N-gram model probability
- Trigram model probability calculation

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$