[illegible] ← generation

[illegible]
[illegible]
[illegible]

cats

→ Surface form vs Lexical form:

Step 1: Identify the morphemes
Step 2: Analyse morphemes.

Orthography changes

చేతి (from చెయ్)
నేత (from నెయ్)

exceptions: వెయ్ → వెయ్
✗ → వేత

Refer: Regularity, analogy in histo-logo.

Machine learning of morph rules

Based on large data, orthographic, suppletion assimilation, vowel harmony etc are learnt, but we need to provide a list of all exceptions which are not regular.

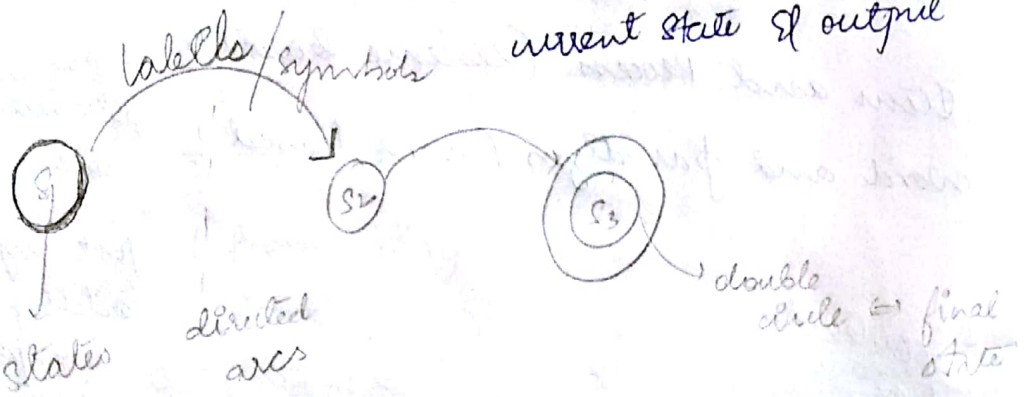
CELEX → lexical database.
↳ lemma, wordform and other details
(headword)

FSA: Machine consisting of

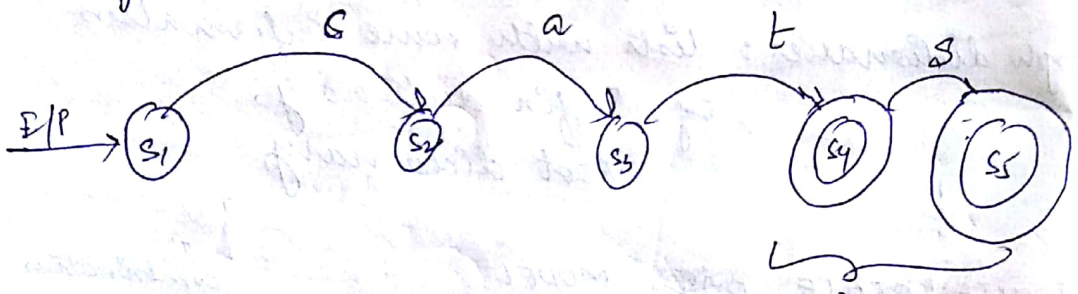
(1) Input tape

(2) Finite number of states, with one initial and one or more accepting states

(3) Actions in terms of transition from one ~~state~~ state to another, depending on current state & output

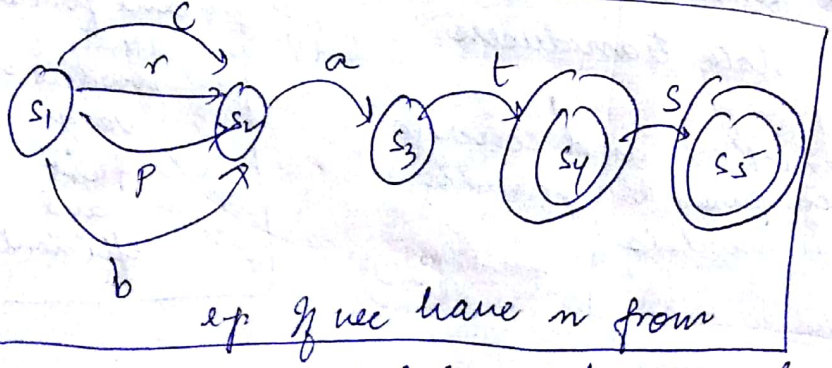


e.g: CATS.



To extend, we do this

This indicates presence of two such words i.e cat & cats



e.g. If we have n from news, it glues S_1 to S_2 and ~~also~~ check for new as root word.

24/9/18

Shallow Parsing (Chunking and LWG)

gives word group and the headword's rel^n with other words.

→ Why morph analysis?

Two levels of information stored in a sentence: ① Lexical level
② Grammar (relative arrangement of words which gives some meaning not explicit in any of the words)

→ We need morph analysis for languages which store a lot of information lexically.
e.g. Dravidian languages.
Whereas for English, we don't require morph analysis, if the main goal is just parsing
e.g. Mohan is leaving for class now.
మోహన్ ఇప్పుడు class కి వెళ్తున్నాడు.

→ No one-one correspondence between form and meanings: leads to ambiguity

→ In some cases, where we need to process a sentence (a little more work than just parsing) we would need morph analysis to extract features such as the roots, suffixes, etc.
e.g. Mohan is leaving for class now.
Mohan will leave for class now.

भा रा भा - Analytical language.
VM VAUX VAUX

ಕನ್ನಡ - Synthetic / Fusional

Inflections

Agglutinative - just concatenates morphemes & words

Synthetic/ Fusional - changes occurs at morpheme boundaries (eg: phonological changes such as vowel harmony)

cf
(राम ने)(मोहन को)(किताब दी)। → Tense, aspect is occuring seperately (not at lexical level)

case breaking acc to pada notion.

राम के हाथ से

"pada" → Since nouns and verbs have to be inflected before they can make a sentence together (either nominal or finite), each inflected form is a pada.

(sanskrit)

THE PROCESS IS CALLED LOCAL WORD GROUPING

singular nominative

singular accusative

27/9/18

- POS tagging, morph analysis and chunking are usually the three main steps followed prior to parsing.

e.g. (My friend [NN])NP (visited [VBD])VP (the new shop [NN])NP (in the outskirts) (of our city [NN])NP.

Chunk Tags for Indian languages

NP		
JJP	adjec. chunk	(बहुत सुन्दर सी)
RBP	adverb chunk	(మెల్ల మెల్లగా)
NEGP	chunk for negatives	(eg: बिना)
CCP	conjunct chunks.	(और is an eg.)
BLK.	miscellaneous ↳ words which can't be grouped under any head.	

- Chunks have to be kept as flat as possible.

Chunk
minimal phrase with internal dependencies not distorted.

WX notation

e ē E (ae) → Telugu
ए ऍ ऐ

eg. (Ram) (aur)CCP (Shyam) → the doubt of which is the head remains

(Ram aur woh ladka)NP

Word compounding and Multiword Expressions (MWE) 4/10/19

- Different conventions are followed.
  eg: whitespaces, hyphens or nothing
  blackboard, blackboard, black-board.

Main elementary problems in NLP:

1. Multiword Expressions (MWE)
2. Word Sense Disambiguation (WSD)
3. Named Entity Recognition (NER)

Main examples of MWE:

1. Compounds (Nominal & Verbal)
2. Named Entities.
3. Idioms
4. (Some common prepositional / post positional / adverbial phrases) ?

NST - A word which is classified as a noun, but could be an adverb or something
eg: NSTs.

मेज़ को ऊपर (vs) मेज़ पर
|
NST

Compounds

Nominal	Verbal	Noun – Verb
Table cover	गिर पड़ना	इंतज़ार करना
Tube light	काट देना	
	ले आना	

- Identifying compounds computationally ?
- MWE – a multiword unit or a collocation of words that co occur together statistically more than chance

↓

Co occurrence is one of the main criteria.

- e.g. जल & पानी are synonyms, but गंगाजल is often used whereas गंगा पानी is not!

  So गंगाजल, strong coffee etc. are compounds (multiword expressions)

Compositional vs Non Compositional

→ black bag

→ white ants (they're termites. Neither white nor ants)

→ kick the bucket (neither kicking or bucket")

★ There is never a binary division of compositional and non compositional

It could be gradual. Even if binary, depends on aspect.

Named E
→ Proper
e.g. ...
(Intern
Placed N

Named Entities

→ Proper nouns

eg How do you POS tag IIIT?

(International Institute of Information Technology)

NP NP NP NP

Khoo