Morphology

Computers and Natural Language Understanding

Introduction

- Most basic unit of linguistics is a word.
- But, words in itself are part of more complex set of more primitive parts.
- Morphology concerns with construction of words from more basic components corresponding roughly to meaning units (Morphemes).
- For Example: Fox contains a single morpheme “Fox” while Foxes contains two morphemes “Fox” and “es”

Morphology

- In this example, we have joined two broad classes of words: stems and affixes.
- The exact details of the distinctions vary from language to language, but intuitively, the stem is the main morpheme of the word.
- Stem supplies the main meaning, while the affixes add additional meanings.

Morphology

- There are two broad ways in which words can be formed using morphemes
  - Inflection :: Combination of word stems with a grammatical morpheme. In this the resulting word is of same class as that of the stem (root) word.
  - Derivation :: Combination of word stems with a grammatical morpheme resulting in a word of a different class as that of the stem (root) word.

Inflectional Examples – Regular Nouns/Verbs

Stem Word	Affix	Resultant
Cat	s	Cats
Box	s	Boxes
Eat	ing	Eating
किताब	ए	किताबें
लडका	ए	लडके

Inflectional Examples – Irregular Nouns

Singular	Plural
Ox	Oxen
Mouse	Mice
Goose	Geese

Inflectional Morphology – Irregular Verbs

Stem	Past	-ed participle
Eat	Ate	Eaten
Catch	Caught	Caught
Cut	Cut	Cut

Derivational Morphology

Base Verb/Adjective	**Suffix**	**Derived Noun**
Computerize (Verb)	ation	Computerization
Appoint (Verb)	ee	Apointee
Kill (Verb)	er	Killer
Fuzzy (Adjective)	ness	Fuzziness

Derivational Morphology

Base Noun/Verb	**Suffix**	**Derived Adjective**
Computation (Noun)	al	Computational
Embrace (Verb)	able	Embraceable
Clue (Noun)	less	Clueless

Morphological Parsing

- Process of separating stems from affixes. For Example

Input	**Morphological Parsed Output**
cats	cat + N + PL
cat	cat + N + SG
cities	city + N + PL
merging	merge + V + Present
caught	catch + V + Past

Morphological Parsing

- Second column contains stems of words with morphological features like noun/verb or singular/plural or part/present etc.
- To build a morphological parser, we require following information:

1. **Lexicon**: list of stems and affix with basic information.

Morphological Parsing

2. **Morphotactics** : Model of morpheme ordering that explains which class of morphemes can follow other classes of morphemes inside a word. It describes the way morphemes are arranged together.
   - For Example – rest – less – ness is a valid English word but not rest – ness – less.

Morphological Parsing

3. **Orthographic Rules** : These are spelling rules that specify the changes that occur when two morphemes combine.
    - **Example –** y -> ier spelling rule changes **easy** to **easier** and not **easyer**

Alternate to Morph Analysis

- Morphological Parsing or Analysis can be avoided if we have an exhaustive list of lexicons that features all word forms of all root words. For example –

Word Form	Category	Root	Gender	Number	Person
पुस्तक	noun	पुस्तक	feminine	Singular	3rd
पुस्तके	noun	पुस्तक	feminine	Plural	3rd
पुस्तको	noun	पुस्तक	feminine	Plural	3rd

Disadvantages of Exhaustive Lexicon

- Heavy demand on memory.
- Fails to show relationship between different roots having similar word forms
- Some languages like Turkish can have infinite word forms. It is impractical to list all possible words for these languages.

Stemmers

- Simplest Morphological Parsing Systems.
- Collapse morphological variation of a given word to one lemma or stem.
- Two widely used stemming algorithms are developed by Lovins(1968) and Porter(1980). There broader working is as follows
  - Suffix removal: Remove predefined endings
  - Recoding: Adds predefined endings to the output of the first step.

Stemmers

- Stemmers are not perfect.
- Problem gets more complex with morphologically rich languages.

Solution

- More efficient two level morphological model were developed in early 1980s.

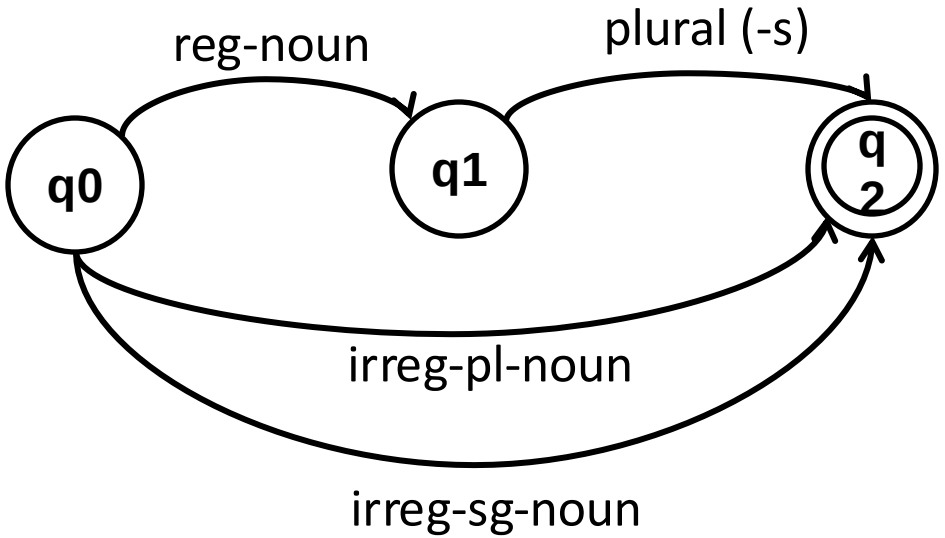
Two Level Model

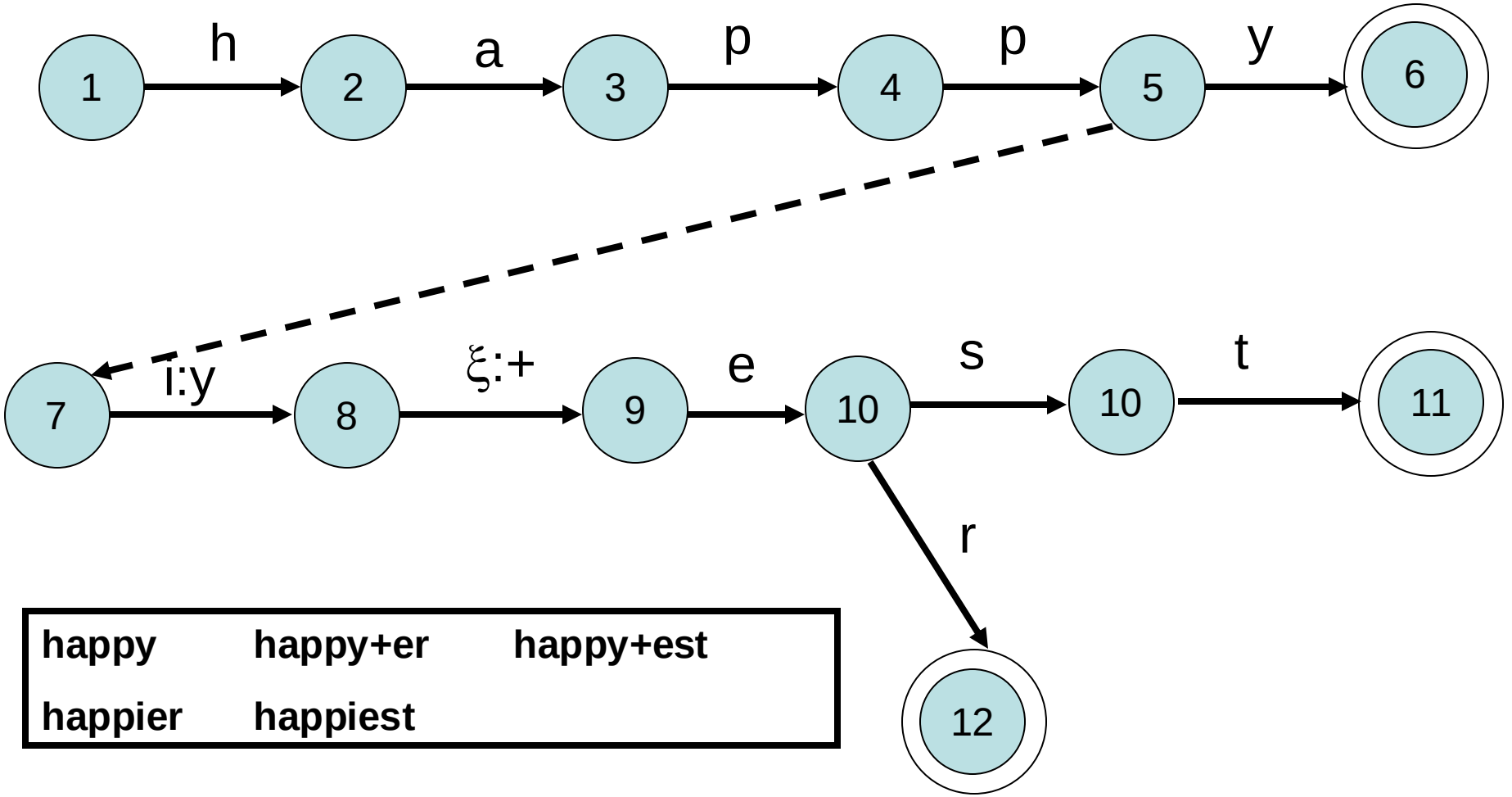
- Word is represented as a correspondence between its lexical level and surface level
- Surface level represents actual spelling of the word.
- Lexical level represents concatenation of its constituent morphemes.

Surface Level	c	a	t	s	
Lexical Level	c	a	t	+N	+PL

Two Level Model

- Implemented with the help of Finite State Automata.
- Popularly known as Finite State Transducers.
- Transducer maps a set of symbols to another.

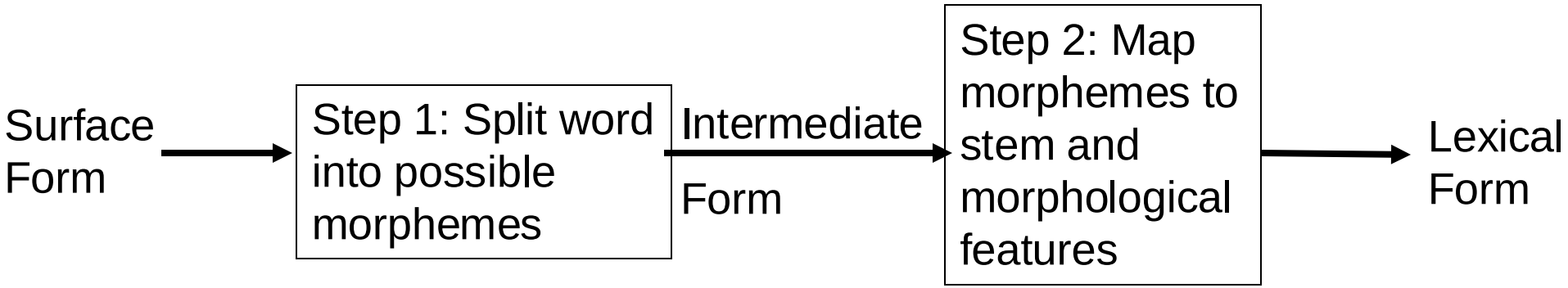




happy	**happy+er**	**happy+est**
happier	**happiest**	

FST Showing different forms of noun ‘happy’

Two Step Morphological Parser



bird	bird	bird + N +sg
birds	bird + s	bird + N + PL
goose	geese	geese + N+PL
geese	geese	geese + N+sg
boxes	box + s	box + N + PL
घोड़े	घोडा + ए	घोडा + N + PL
पुस्तक	पुस्तक	पुस्तक + N + sg
पुस्तके	पुस्तक + ए	पुस्तक + N + PL
रोटी	रोटी	रोटी + N + sg

Application of Morph Analyzer

- **Spell Checker** – Find errors and corrections.
- **Information Retrieval** – Find different occurrence of root word on WebPages.
- **Natural Language Parsing** – Find words with associated grammatical information.