CS565: INTELLIGENT SYSTEMS AND INTERFACES

Text Normalization

Semester: July – November 2020

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Recap

- Corpora

- Text Normalization: Tokenization

Objective

- Text Normalization
  - Word Normalization
  - Sentence Segmentation

TEXT NORMALIZATION

Word Normalization

Definition

- Converting the words in a standard format, i.e. choosing a single canonical form for words which can appear in multiple forms. Example: Ph.D., PhD., PhD,

Word Normalization: Case Folding

- Conversion into lowercase
- May be good idea for Information Retrieval (search) purpose
- May not be good for POS tagging or NER (US: the country VS us: pronoun)

What happens to cases like: “*the*”, “*The”, and “THE”* vs. “*Mr. Brown*” and “*brown paints*”

Word Normalization: Lemmatization

- Task of determining two words have the same root, same POS, same sense but may have different word forms.
- Mostly relevant for IR (search) purpose
- Example: *I am learning -> I be learn*
- *Stems:* supplying the main meaning
- *Affixes:* supplying the additional meaning
- Requires **Morphological Parsing** of words

Stemming

- Crude form of lemmatization
- Consists of chopping off word-final affixes
- Done with a series of rewrite rules applied consequently one after the other, i.e., in a cascade.
- Sample of such rules
  - *Ational -> ATE (relational -> relate)*
  - *SSES -> SS (grasses -> grass)*
- Porter Stemmer

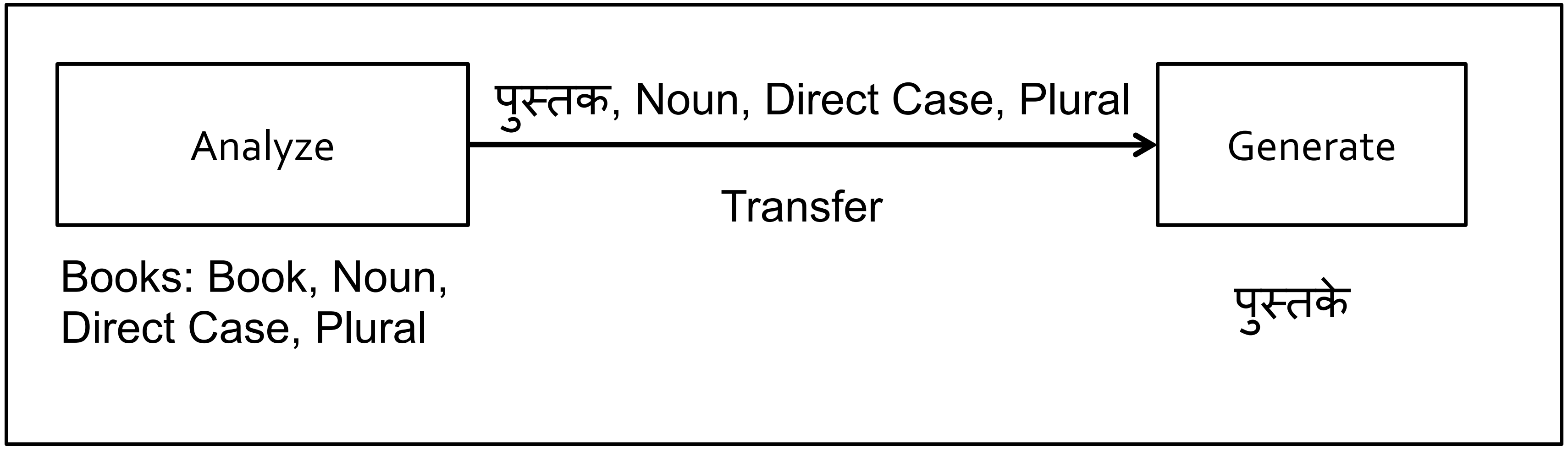
Morphology: Exploring structure of words

- Words have structure
  - *Foxes* breaks down into *Fox* and *–es*
  - *Unknowingly* is derived from *knowingly*, which is derived from *knowing*, which in turn is derived from *know*
- Morphology: study of *minimal meaning bearing*, referred to as *morphemes*, units in words
  - *Fox* and *–es* in *Foxes*
  - *Un, know, -ing, -ly* in *Unknowingly*

Why study of Morphology matters ?

- Information Retrieval (Stemming)
  - Useful to map all of *learning*, *learns*, *learned* to *learn*

- Machine Translation



Adapted from IIT Bombay lecture slides

Why study of Morphology matters ?

- Efficiency
  - Cannot list all possible forms even in morphological poor language (relatively) English
  - *Productivity* of language

- Morphological rich languages
  - Turkish, Finnish, Indian Languages

Two types of Morphemes

- Stems
  - Main morpheme of the word, supplying the main meaning
  - Example: fox, know

- Affixes
  - Provides additional meanings of various kinds
  - Mainly categorized into four types -
    - Prefix: Un-, Im-
    - Suffix: -s, -es, -ly
    - Infix: Mostly with other language. –*n*- in “vandimi” in Sanskrit; -um- in humingi in Philipine language Tagalog
    - Circumfix: ge-sag-t (meaning: said) in German; past participle of the verb *sagen* (to say)

Concatenative and non-concatenative Morphology

- Concatenative
  - Word is composed by concatenating a number of morphemes
  - Prefixes and Suffixes
- Non-concatenative
  - Combining morphemes is more complex
  - Tagalog Infixation example (hingi + um -> humingi)
  - Templatic morphology
    - Arabic, Hebrew
    - Hebrew: verb constitutes a root (carrying basic meaning) and a template giving ordering of consonants and vowels determining semantic information (active, passive)
    - Example: lmd (learn or study), template: CaCaC -> lamad (he studied)
    - Example: lmd (learn or study), template: CuCaC -> lumad (he was taught)

Two broad classes of Morphology

- Inflection
  - Stem + grammatical morpheme (s)
  - Usually word of the same class and filling some syntactic functions
  - English has simple inflectional morphology, compared to Hindi, Finnish or other European Languages
  - Very productive
- Derivation
  - Stem + grammatical morpheme (s)
  - Usually results in a word of different class and often difficult to guess exact meaning
  - English also has quite complex derivational morphology
  - Relatively less productive (-ation cannot be added to all verbs)

Inflectional Morphology: Example

- Nouns
  - Suffixes for Plural and possessive
- Verbs
  - Suffixes for –s form, -ing participle, past form or –ed participle
  - Walks, walking, walked
- Adjectives
  - Suffixes for comparatives
  - Cheap, cheaper, cheapest

Derivational Morphology: Example

- Nominalization
  - Formation of new nouns, often from verbs or adjectives
  - Organize (v) + -ation
  - Appoint (v) + -ee
  - Silly (ADJ) + -ness
- Adjectives
  - Computation (N) + -al

Derivational Morphology: Example

- Nominalization
  - Formation of new nouns, often from verbs or adjectives
  - Organize (v) + -ation
  - Appoint (v) + -ee
  - Silly (ADJ) + -ness

- Adjectives
  - Computation (N) + -al

Morphological Analysis

- Token -> stem + POS +grammatical features
  - Cats -> Cat +N +PL
- Often non-deterministic
  - Plays -> play +N +PL
  - Plays -> play +V +3SG
- Lemmatization
  - Token -> stem

Parsing the morphological structure

- Goal
  - Given an input word in *surface form*, produce *stem* plus *morphological features* (POS and grammatical features) as an output

- Example Goal: Productive nominal plural (-s) and the verbal progressive (-ing)
  - Input: Cats ; Output: cat +N +PL
  - Input: Eating; Output: eat +V +PRES-PART

Three knowledge resources needed

- Lexicon
  - Repository of words in a language
  - Explicit list is infeasible. Why ?
  - List of stems and affixes with basic information about them

- Morphotactics
  - Rules of morpheme ordering
  - Example: English plural morpheme follows the noun rather than preceding it.

- Orthographic or Spelling Rules
  - Model change in spelling when two morphemes combine
  - Fly -> flies [y -> ie]

Morphological Analyzer

- FSAs can be used for morphological recognizers
- Morphological analyser produce output
  - Input: cats
  - Output: cat +N +PL
- Finite state Transducer to model two level morphology
  - Lexical level: concatenation of morphemes
  - Surface level: actual spelling of the word
  - Alphabets of complex symbols

Lexicons and Morphotactics

- Structured as
  - List of stems and affixes
  - Representation of the syntactics of morphemes

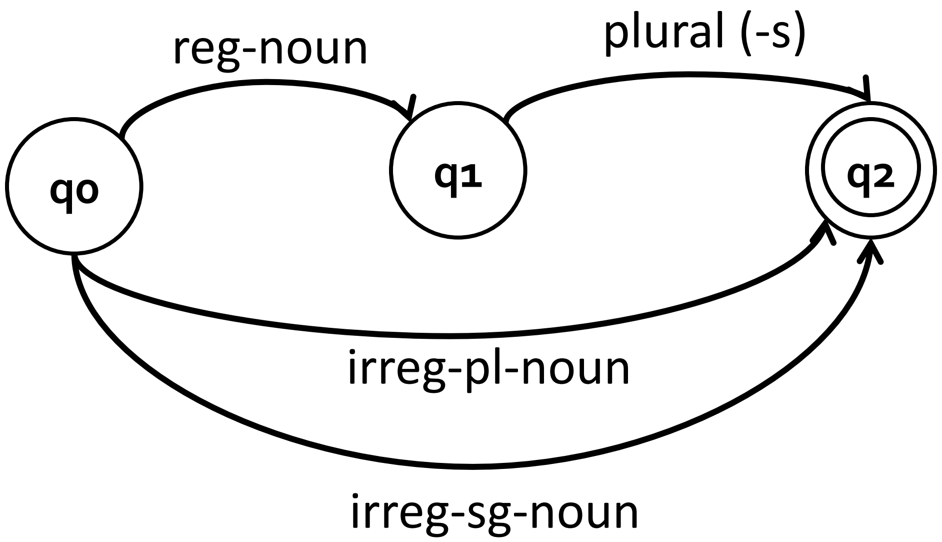
- Represent via a finite-state automaton (FSA)

Example: A FSA for English nominal inflection

Takes regular nouns (reg-nouns) that take regular –s plural.

Also includes irregular noun forms, both singular and plural, that don't take –s

Ignore mistakes like foxes.



Current Status

- Learning from data
  - Unsupervised and supervised parsing

- Good Resource
  - SIGMORPHON workshop and associated challenges

TEXT NORMALIZATION

Sentence Segmentation

Defining Sentence Boundary

- Something ending with a '.', '?', or '!'
  - Language specific
- Problem with '.'
  - Still 90% of periods are sentence boundary indicators [Riley 1989].
- Sub-sentence structure with the use of other punctuation
  - "The scene is written with a combination of unbridled passion and sure-handed control: In the exchanges .... inexorability of separation"
- Other issues
  - "You remind me," she remarked, "of your mother."

Defining Sentence Boundary: A heuristic

- Put putative sentence boundaries after occurrences of ., ?, ! (and may be ;, :, -)
- Check presence of following quotation marks, if any move the boundary.
  - "You remind me," she remarked, "of your mother."
- Disqualify a period boundary if –
  - It is preceded by a known abbreviation that does not generally occur at the end of sentence such as Dr., Mr. or vs.
  - It is preceded by a know abbrev. that is generally not followed by an uppercase word such as etc. or Jr.
- Disqualify a boundary with a ? or ! If
  - It is followed by a lowercase letter (or name)

Issues with Heuristic or set of pre-defined rules

- Is it possible to define such rules without the help of experts?
- Will it work for all languages?

Machine Learning Methods: Sentence boundary as classification problem

- Riley (1989) used classification trees
  - Features: case & length of the words preceding and following a period; prior prob of words occurring before and after a sentence boundary etc.
- Palmer and Hearst (1997) used neural network model
  - Instead of prior probability, PoS distribution of the preceding and following words.
  - Language-independent model with accuracy of 98-99%
- Reynar and Ratnaparkhi (1997) and Mikheev (1998) used Max. Ent approach
  - Language independent model with accuracy of 99.25%

References

- Chapter 4 [FSNLP]
- Chapter 2 [Jurafsky and Martin 3rd Ed.]