CSE528

Natural Language Processing

Venue:ADB-405 SLOTS: A2+TA2 Topic: Morphology

Prof. Tulasi Prasad Sariki,

SCSE, VIT Chennai Campus

www.learnersdesk.weebly.com

Contents

- What is Morphology
- Where Morphology is useful?
- Classification of Morphemes
- Properties of affixes
- Representation / Terminology
- Example

What is Morphology?

Morpheme is a minimal meaning-bearing unit in a language.

Morphemes are abstract concepts denoting entities or relationships.

Morphology is the study of the internal structure of words.

In natural languages, words are made up of meaningful subunits called morphemes.

Morphological parsing is the task of recognizing the morphemes inside a word

e.g., *hands, foxes, children*

Where Morphology is useful?

Machine translation

Information retrieval

Lexicography

Any further processing (e.g., part-of-speech tagging)

Observations about words and their structure

Some words can be divided into parts which still have meaning

Many words have meaning by themselves. But some words have meaning only when used with other words

Some of the parts into which words can be divided can stand alone as words. But others cannot

These word-parts that can occur only in combination must be combined in the correct way

Languages create new words systematically

Classification of Morphemes

Lexical morphemes are those that having meaning by themselves (more accurately, they have **sense**).

Nouns, verbs, adjectives ({boy}, {buy}, {big}) are typical lexical morphemes.

Grammatical morphemes specify a relationship between other morphemes. But the distinction is not all that well defined.

Prepositions, articles, conjunctions ({of}, {the}, {but}) are grammatical morphemes.

Classification of Morphemes

Free morphemes – morphemes which can stand by themselves as separate words,

e.g. *structure*, *like*, go, work, friend etc.

Bound morphemes – morphemes which cannot normally stand alone but need to be attached to other forms,

e.g. re-, -ed, -s, -ing etc.

- unit of meaning which can only exist alongside a free morpheme.
- Bound morphemes operates in the connection processes by means of ***derivation, inflection***, and ***compounding***.

Classification of Morphemes

We can usefully divide morphemes into two classes

- **Root or Lexeme**: The core meaning-bearing units
- **Affixes**: Bits and pieces that adhere to stems to change their meanings and grammatical functions
  - Prefix: un-, anti-, etc
  - Suffix: -ity, -ation, etc
  - Infix: are inserted inside the stem, English has almost no true infixes
  - Circumfixes – a discontinuous morph composed of two parts which embrace the base element (live $\rightarrow$ en-live-en $\rightarrow$ enliven)

Properties of roots

- Main part of word
- Must be at least one in a word
- In English, limited to two in a word
  - (simple words have one, compound words have two);
- Can occur independently
- Tend to have richer, more specific semantic content
- Position is relatively free with respect to other roots
  - E.g. photograph vs. telephoto

Properties of affixes

- Subordinate part of word
- Not necessarily present--some words occur without any
- Multiple affixes can occur in a word
- Are dependent (bound) elements
- Have more "schematic" (non-specific) content; often grammar-like function
- Can either precede or follow their roots (prefixes and suffixes ,respectively)
- Position for a given affix with respect to root is fixed

Example

Given word: Unbreakable

How many morphemes?

comprises **three** morphemes

un- (a bound morpheme signifying "not")

-***break***- (the root, a free morpheme)

-***able*** (a free morpheme signifying "can be done").

Representation / Terminology

Morphological: *girls* = {girl} + {s}

Semantic: {girl} = [-adult; -male; +human, ...] + {s} = {PLU} = [plural]

Braces, { } indicate a morpheme. Square brackets, [ ]

indicate a semantic characterization. Italics indicate a lexical item.

Two different morphemes may be pronounced the same way.

Ex: *–er in buyer* and *shorter*

Morphemes and Words

Combine morphemes to create words.

Inflectional Morphology

Combination of stem and morpheme resulting in word of same class Usually fills a syntactic feature such as agreement

E.g., plural –s, past tense -ed

Derivational Morphology

Combination of stem and morpheme usually results in a word of a different class Meaning of the new word may be hard to predict

E.g., +ation in words such as computerization

Inflectional Morphology

Inflection is a morphological process that adapts existing words so that they function effectively in sentences without changing the category of the base morpheme.

Word stem + grammatical morpheme cat + s

Only for nouns, verbs, and some adjectives

Nouns

- plural:

Rules for regular: +s, +es irregular: mouse-mice; ox-oxen

Rules for exceptions: e.g.-y -> -ies like: butterfly-butterflies

Inflectional Morphology (verbs)

Morphological form	Regular Inflected form			
stem	walk	thrash	try	map
-s form	walks	thrashes	tries	maps
-ing form	walking	thrashing	trying	mapping
-ed form(past)	walked	thrashed	tried	mapped

Inflectional Morphology (verbs)

Morphological form	Irregular Inflected form		
stem	eat	catch	cut
-s form	eats	catches	cuts
-ing form	eating	catching	cutting
-ed form(past)	ate	caught	cut
-ed form(participle)	eaten	caught	cut

Inflectional Morphology (verbs)

The suffix –s functions in the Present Simple as the third person marking of the verb

- to work – he work-s

The suffix –ed functions in the past simple as the past tense marker in regular verbs

- to love – lov-ed

The suffixes –ed (regular verbs) and –en (for some regular verbs) function in the marking of the past partciple

- to study studied / To eat eaten

The suffix –ing functions in the marking of the present participle.

- to eat – eating / To study - studying

Inflectional Morphology (nouns)

Regular Nouns (cat, hand) Irregular Nouns(child, ox)

Morphological form	Regular Inflected form	
stem	cat	hand
-s form(plural)	cats	hands
Morphological form	Irregular Inflected form	
stem	child	ox
-s form(plural)	children	oxen

The suffix –s functions in the marking of the plural of nouns: dog – dogs

The suffix –s functions as a possessive marker: Laura – Laura's book.

Regular vs Irregular

It is a little complicated by the fact that some words misbehave (refuse to follow the rules)

- Mouse/mice, goose/geese, ox/oxen
- Go/went, fly/flew

The terms regular and irregular are used to refer to words that follow the rules and those that don't.

Inflectional Morphology (Adjectives)

The suffix –er functions as comparative marker: quick – quicker

The suffix –est functions as superlative marker: quick - quickest

Derivational Morphology

Derivation is concerned with the way morphemes are connected to existing lexical forms as **affixes**.

We distinguish affixes in two principal types

- **Prefixes** - attached at the beginning of a lexical item or base-morpheme – ex: un-, pre-, post-, dis, im-, etc.
- **Suffixes** – attached at the end of a lexical item

– ex: -age, -ing, -ful, -able, -ness, -hood, -ly, etc.

Examples of Derivational Morphology

Lexical item (free morpheme): **like** (verb)+ prefix (bound morpheme) **dis-**= dislike (verb);

like + suffix **–able** = likeable + prefix **un-** =unlikeable + suffix **–ness** = unlikeableness

like + prefix **un-** = unlike + suffix **–ness** = unlikeness

like + suffix **–ly** = likely + suffix **–hood** =likelihood + prefix **un-** =unlikelihood

Derivational Morphology

Derivational affixes can cause semantic change

Prefix **pre-** means *before;* **post-** means *after*; **un-** means *not*, **re-** means *again.*

Prefix = fixed *before*; Unhappy = *not* happy = sad; Retell = tell *again*.

Prefix **de-** added to a verb conveys a sense of subtraction; **dis-** and **un-** have a sense of negativity.

to decompose; to defame; to uncover; to discover.

Derivational Morphology

Derivational affixes can mark category change

For Nouns

Suffix	Base Verb / Adjective	Derived Noun
-ation	Computerize (V)	Computerization
-ee	Appoint (V)	Appointee
-er	Kill (V)	Killer
-ness	Fuzzy (A)	Fuzziness

For Adjectives

Suffix	Base Verb / Noun	Derived Adjective
-al	Computation (N)	Computational
-able	Embrace (V)	Embraceable
-less	Care (N)	Careless
-ful	Care (N)	Careful

Derivational Morphology

Verb Clitics are usually weak forms of functional elements

Full Form	Clitic
am	’m
is	‘s
are	‘re
will	‘ll
have	‘ve
has	‘s
had	‘d
would	‘d

Derivational Processes

1. Derivation: (or Derivational affixation, Affixation)

   antiintellectualism

2. Compounding: combine two or more morphemes to form new words

   bathroom, blackboard

3. Reduplication: full or partial repetition of a morpheme

   dilly-dally, zig-zag

4. Blending: parts of the words that are combined are deleted

   fantastic + fabulous -> fantabulous

Derivational Processes

5. Clipping: part of a word has been clipped off

Prof , lab, doc

6. Acronyms: abbreviate a longer term by taking the initial letters

WHO -> World Health Organization

7. Back formation: A word (usually a noun) is reduced to form another word of a different type (usually a verb)

television -> televise

babysitter -> babysit

Derivational Processes

8. Extension of word formation rules : Part of a word is treated as a morpheme though it's not

    workaholic

9. Functional shift (Conversion): A change in the part of speech

    computer users today use a mouse and bookmark an Internet address

10. Proper names -> Common words

    Xerox -> Photo copy

    JCB -> Proclainer

Derivational Processes

11. Coining: Creating a completely new free morpheme

    googol -> 10^{100}

12. Onomatopoeia: words imitate sounds in nature

    tick-tock, quack

13. Borrowing: The taking over of words from other languages French to English

    brigade, ballet, bigot

Derivational Processes

Many paths are possible.

Start with compute

Computer -> computerize -> computerization

Computer -> computerize -> computerizable

Computation -> computational

But not all paths/operations are equally good (allowable?)

Clue

Clue -> *clueable

Happy→ unhappy

Sad → *unsad

Derivational Processes

Morphotactics

Morphotactics is concerned with ordering of morphemes.

The ordering restrictions in place on the ordering of morphemes

antiintellectualism -anti -ism -al -intellect anti + intellect + al +ism

Morphophonemics:

Focus on the sound changes that take place in morphemes when they combine to form words.

e.g., the vowel changes in “sleep” and “slept,” “bind” and “bound,” “vain” and “vanity,” and the consonant alternations in “knife” and “knives,”.

Derivational Processes

Semantics: In English, un- cannot attach to adjectives that already have a negative connotation:

Unhappy vs. *unsad Unhealthy vs. *unsick Unclean vs. *undirty

Phonology: In English, -er cannot attach to words of more than two syllables

great, greater Happy, happier

Competent, *competenter Elegant, *eleganter

Inflectional vs Derivational

	Inflectional	Derivational
Lexical category	Do not change the lexical category of the word.	Often change the lexical category of the word
Location	Tend to occur outside derivational affixes.	Tend to occur next to the root
Type of meaning	Contribute syntactically conditioned information, such as number, gender, or aspect.	Contribute lexical meaning
Affixes used	Occur with all or most members of a class of stems.	Are restricted to some, but not all members of a class of stems
Productivity	May be used to coin new words of the same type.	May eventually lose their meaning and usually cannot be used to coin new terms
Grounding	Create forms that are fully-grounded and able to be integrated into discourse.	Create forms that are not necessarily fully grounded and may require inflectional operations before they can be integrated into discourse

Stemming

Stemming

Stemming algorithms strip off word affixes yield stem only, no additional information (like plural, 3rdperson etc.) used, e.g. in web search engines.

Stemming is one technique to provide ways of finding morphological variants of search terms.

Used to improve retrieval effectiveness and to reduce the size of indexing files.

Reduce tokens to “root” form of words to recognize morphological variation.

“computer”, “computational”, “computation” all reduced to same token “compute”

Stemming

Criteria for judging stemmers

- Correctness
  - Overstemming: too much of a term is removed.
  - Understemming: too little of a term is removed.
- Retrieval effectiveness
  - Measured with recall and precision, and on their speed, size, and so on
- Compression performance

Type of stemming algorithms

Table lookup approach

Successor Variety

n-gram stemmers

Affix Removal Stemmers

Table lookup approach

Store a table of all index terms and their stems, so terms from queries and indexes could be stemmed very fast.

Problems

- There is no such data for English. Or some terms are domain dependent.
- The storage overhead for such a table, though trading size for time is sometimes warranted.

Successor Variety

Determine word and morpheme boundaries based on the distribution of phonemes in a large body of utterances.

The successor variety of a string is the number of different characters that follow it in words in some body of text.

The successor variety of substrings of a term will decrease as more characters are added until a segment boundary is reached.

Successor Variety Example

Test Word: ***READABLE***

Corpus: ABLE, APE, BEATABLE, FIXABLE, READ, READABLE, READING, READS, RED, ROPE, RIPE

Prefix	Successor Variety	Letters
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	(Blank)

Successor Variety Example

cutoff method

- some cutoff value is selected and a boundary is identified whenever the cutoff value is reached

peak and plateau method

- segment break is made after a character whose successor variety exceeds that of the characters immediately preceding and following it

complete method

entropy method

Successor Variety

Two criteria used to evaluate various segmentation methods

1. the number of correct segment cuts divided by the total number of cuts
2. the number of correct segment cuts divided by the total number of true boundaries

After segmenting, if the first segment occurs in more than 12 words in the corpus, it is probably a prefix.

The successor variety stemming process has three parts

1. determine the successor varieties for a word
2. segment the word using one of the methods
3. select one of the segments as the stem

n-gram stemmers

Association measures are calculated between pairs of terms based on shared unique digrams.

```
statistics => st ta at ti is st ti ic cs
unique digrams = at cs ic is st ta ti
statistical => st ta at ti is st ti ic ca al
unique digrams = al at ca ic is st ta ti
```

Dice's coefficient (similarity)

$$S = \frac{2C}{A+B} = \frac{2*6}{7+8} = .80$$

A and B are the numbers of unique digrams in the first and the second words. C is the number of unique digrams shared by A and B.

n-gram stemmers

Similarity measures are determined for all pairs of terms in the database, forming a similarity matrix

Once such a similarity matrix is available, terms are clustered using a single link clustering method

Affix Removal Stemmers

Affix removal algorithms remove suffixes and/or prefixes from terms leaving a stem

- If a word ends in “ies” but not ”eies” or ”aies ” **(Harman 1991)**

  Then “ies” -> “y”
- If a word ends in “es” but not ”aes” , or ”ees ” or “oes”

  Then “es” -> “e”
- If a word ends in “s” but not ”us” or ”ss ”

  Then “s” -> “NULL”

The Porter Stemmer

Online Demo: http://9ol.es/porter_js_demo.html

Typical rules in Porter stemmer

sses→ss (caresses -> caress)

ies→i (ponies -> poni, ties -> ti)

ational→ate

tional→tion

ing →ε (motoring -> motor)

Conditions on the stem

1. The measure , denoted m ,of a stem is based on its alternate vowel-consonant sequences.

$[C](VC)^m[V]$ Square brackets indicate an optional occurrence.

Measure	Example
M=0	TR,EE,TREE,Y,BY
M=1	TROUBLE,OATS,TREES,IVY
M=2	TROUBLES,PRIVATE,OATEN

E.g.,

Troubles

C V CVC

Conditions on the stem

2.*<X> ---the stem ends with a given letter X

3.*v*---the stem contains a vowel

4.*d ---the stem ends in double consonant

5.*o ---the stem ends with a consonant-vowel-consonant, sequence, where the final consonant is not w, x or y

6. *s --- the stem ends with a given letter S

Step1

SSES -> SS

caresses -> caress

IES -> I

ponies -> poni

ties -> ti

SS -> SS

caress -> caress

S -> ϵ

cats -> cat

Step2a

(m>1) EED -> EE

Condition verified: agreed -> agree Condition not verified: feed -> feed

(*V*) ED -> ϵ

Condition verified: plastered -> plaster Condition not verified: bled -> bled

(*V*) ING -> ϵ

Condition verified: motoring -> motor Condition not verified: sing -> sing

Step2b

(These rules are ran if second or third rule in 2a apply)

AT-> ATE	conflat(ed) -> conflate
BL -> BLE	Troubl(ing) -> trouble

(*d & ! (*L or *S or *Z)) -> single letter

Condition verified: hopp(ing) -> hop, Condition not verified: fall(ing) -> fall

(m=1 & *o) -> E

Condition verified: fil(ing) -> file Condition not verified: fail -> fail

Steps 3 and 4

Step 3: Y Elimination (*V*) Y -> I

Condition verified: happy -> happi

Condition not verified: sky -> sky

Step 4: Derivational Morphology, I

(m>0) ATIONAL -> ATE

Relational -> relate

(m>0) BILITI -> BLE

sensibiliti -> sensible

(m>0) IZATION -> IZE

generalization-> generalize

Steps 5 and 6

Step 5: Derivational Morphology, II

(m>0) ICATE -> IC

triplicate -> triplic

(m>0) FUL -> ϵ

hopeful -> hope

(m>0) NESS -> ϵ

goodness -> good

Step 6: Derivational Morphology, III

(m>0) ANCE -> ϵ

allowance-> allow

(m>0) ENT -> ϵ

dependent-> depend

(m>0) IVE -> ϵ

effective -> effect

Step7 (cleanup)

Step 7a

(m>1) E -> ϵ

probate -> probat

(m=1 & !*o) NESS -> ϵ

goodness -> good

Step 7b

(m>1 & *d & *L) -> single letter

Condition verified: controll -> control Condition not verified: roll -> roll

END