Morphology

Why are we Studying Morphology?

- The knowledge of words will help us process language computationally at word level.
- Knowledge of words include word structure and word formation rules.
- This knowledge will help us in developing NLP tools like "Morphological Analyzers" and "Morphological Generators"
- These are important in Information Retrieval, Machine Translation, Spell checking, etc.

What's Morphology?

- The study of word structure
- The study of the mental dictionary:
  - How are words stored in the mind?
  - What is a possible word?
- Example:

  (i) When their mother signaled, the girls barried home unhappily.

The word ‘barried’ do not exist. However, assuming they are valid words of English, we ‘guess’ the meaning by context and the position of the word in the given sentence. We do this using our general knowledge and linguistic knowledge.

When their mother signaled, the girls barried home unhappily.

Part of Speech = verb [position]

[ends in –ed] = past tense

Base form = barry

Meaning = go

The word form is more like carried .

Morphemes

- Have a sound [form] and a meaning:

Example: “cats”

- /kaet/ “four-legged animal”
- /-s/ “plural number”

- Even though /-s/ has a sound and a meaning, it can’t mean “plural” by itself.
- It has to attach to a noun.

"A morpheme is the smallest unit of word form that has meaning"

Examples:

cats = cat + -s

girlish = girl + -ish

unfriendly = un- + friend + -ly

cat, -s, girl, -ish, un-, friend, -ly are morphemes

Function words

Function words:

- Are the words used to make the sentence grammatical
- They have little lexical meaning
- Belongs to closed class category
- Used mainly for determining structure of the sentence

Examples:

Determiners , pronouns, prepositions, auxiliary verbs, conjunctions, articles, etc

Content words/Lexical words

Content words:

- Are the words used to convey what are the important concepts in the sentence
- They have strong lexical meaning
- They are nouns, verbs, adjectives, adverbs, etc
- Used mainly for determining topic of the sentence

Examples:

nouns, verbs, adjectives, adverbs, etc

Morpheme Types

- Free morphemes vs. Bound morphemes
- Lexical morphemes vs. Functional morphemes
- Null/Zero morpheme
- Inflectional morphemes vs. Derivational morphemes
- Root morphemes vs. Affix morphemes

Free vs Bound morphemes

- *electr-* and *tox-* have isolable meanings in *electric, electrify, toxic, (de-)toxify*
- But they cannot be pronounced on their own: they are bound morphemes
- *girl* and *book* have isolable meanings in *girls, girlish, books, booked, booking*
- They can occur on their own: they are free morphemes

Lexical morpheme

free morphemes: apple, smart, book, slow, eat, write

They can exist on their own as independent words.

bound morphemes: -ceive, -ject, cran-, -ship, un-, dis-

They cannot be used independently.

They need another morpheme [free or bound] to form a word.

Eg: re-ceive, con-ceive, sub-ject, pro-ject, cran-berry, scholar-ship, fellow-ship, un-kind, dis-obey

Functional morphemes

free morphemes:

of, with, she, it, and, although, however, because, then

bound morphemes:

-s, -ed, -ing, -er, -est

Forming words from Morphemes: Inflection and Derivation

- ***Inflectional morphology*** is the process by which a root form of a word is modified by adding prefixes or suffixes that specify its grammatical function but do not changes its part-of-speech.
- We say that a lemma (root form) is inflected (modified/combined) with one or more morphological features to create a surface form.

Forming words from Morphemes: Inflection and Derivation

- ***Derivational morphology*** is the process by which a root form of a word is modified by adding prefixes or suffixes that specify its grammatical function resulting in change in its part-of-speech.
- We say that a *lemma* (root form) is ***derived*** with one or more *morphological features* to create a surface form.

Inflection vs. Derivation

- **Derivational affixes:** allow us to make new words that alters the meaning.
  - *There is an error in the computation. [compute$_V$ – computation$_N$]* {Nominalisation}
  - *It is a computational approach. [computation$_N$ – computational$_{Adj}$]* {Adjectivization}
- **Inflectional suffixes**: required in order to make the sentence grammatical

  Inflected words belong to the same class
  - **Yesterday I walk to class [walk$_V$ – walked$_V$]*
  - **I like all my student [student$_N$ – students$_N$]*

Inflectional Morphology

Examples: [the POS remains the same]

VERBS

EAT = eat, eats, ate, eaten, eating

DRINK = drink, drinks, drank, drunk, drinking

PLAY = play, plays, played, played, playing

0, -s, -ed, -en, -ing are inflectional morphemes

NOUNS

PLAY = play, plays

GIRL = girl, girls

SHEEP = sheep, sheep

-s, 0 are inflectional morphemes

Derivational Morphology

Two types:

- May change the category {N,V,A,Adv}

  $drive_V$ +er = $driver_N$

  eat_v + able = $eatable_{adj}$

  $girl_{N\,+}$ ish = $girlish_{adj}$

  $disturb_{V\,+}$ ance = $disturbance_N$

- Doesn't have to change the category

  un + do_V = $undo_V$

  re+fry_v = $refry_v$

  un+$happy_{Adj}$ = $unhappy_{Adj}$

Derivational – more examples

Verbs

eat – eatable [adj], eatables [noun]

drink – drinking [noun]

play – player [noun]

-able, -ing, -er are derivational morphemes

Nouns

play – playful [adj], replay [verb]

girl – girlish [adj], girlhood [noun]

sheep – sheepish[adj]

-ful, re-, -ish, -hood are derivational morphemes

Four-way Contrasts

- **Lexical, Free:** Nouns, Verbs, Adj, Adv

  cat, town, call, house, hall, smart, fast
- **Lexical, Bound:** including derivational affixes

  rasp- [raspberry], cran- [cranberry] , -ceive [conceive, receive], un-, re-, pre-
- **Functional, Free**: Prepositions, Articles, Conj

  with, at, and, an, the, because
- **Functional, Bound**: inflectional affixes

  -s, -ed, -ing [eats, walked, laughing]

Exercise 1: Identify the free and bound morphemes in the following words

- walked, talked, danced, arrived
- playhouse, watchdog, football player
- drinking, playing, eating
- import, export, transport
- raspberry, cranberry
- invert, convert, divert
- books, pens, boards
- writer, caretaker, rider, fighter

Can the following words be decomposed?
delight, news, traitor, bed, evening

Exercise 2: Identify the lexical and functional morphemes in the following words. Mention if they are free or bound.

- politically
- beautiful
- between
- writing
- raspberries
- unable
- nationalization

Morphology Processing and Morphology Generation

Morphology Processing or Morphology Analysis

- The word **banks** came from the root word **bank**.
- Take the word **banks** and split it into 2 pieces :

  bank + s i.e.,

  (root + affixes)

This process is known as morphology processing or morphology analysis.

Morphology Generation or Synthesis

- We have the root word and from the root word we should be able to produce the word forms.

Example:

nation – nation**al** – national**ism**-national**ity**

Compounding

Compounds made of more than one stem:

- armchair, family man, milkman
- olive oil, sunflower oil, baby oil
- almond biscuits, Osmania biscuits, dog biscuits
- stir-fried rice, bottle-opener
- white collar, blue collar
- dashaanan, pitaambar

Null/Zero morpheme

- A null morpheme is a morpheme that is realized by a phonologically null affix (an empty string of phonological segments)
- A null morpheme is an "invisible" affix.
- It's also called zero morpheme; the process of adding a null morpheme is called null affixation.

Examples

- *cat* = *cat* + *-0* = ROOT("cat") + SINGULAR
- *cats* = *cat* + *-s* = ROOT("cat") + PLURAL

- sheep = sheep + -0 = ROOT(“sheep") + *SINGULAR*
- sheep = sheep + *-0* = ROOT(“sheep") + PLURAL

More examples

- darken[verb] = dark [adj] + -en

redden [verb] = red + -en [make more Red]

yellow [verb] = yellow + 0

brown [verb] = brown + 0

blacken [verb] = black + -en

Root Morphemes vs Affix morphemes

Root morphemes are morphemes around which larger words are built.

Root morphemes are free or bound.

Affixes are additional morphemes added to roots to create multi- or poly-morphemic words.

Affixes are always bound.

- Rats
  - Root = rat [free morpheme]
  - Affix = -s [bound morpheme]
- Project
  - Root = -ject [bound morpheme]
  - Affix = pro- [bound morpheme]
- Mice
  - Root = mouse [free morpheme]
  - Affix = -s [bound morpheme]
- Ate
  - Root = eat [free morpheme]
  - Affix = -ed [bound morpheme]

- Disgracefulness
  - Root = grace [free morpheme]
  - Affixes = dis-, -ful, -ness [bound morpheme]

Affixes

- Morphemes added to free forms to make other free forms are called affixes.
- Mainly four kinds of affixes:
  1. Prefixes (at beginning) – “un-” in “unable”
  2. Suffixes (at end) – “-ed” in “walked”
  3. Circumfixes (at both ends) – “en—en” in enlighten
  4. Infixes (in the middle) – “-um-” in kumilad [‘to be red’], fumikas [‘to be strong’]

     [kilad = ‘red’, fikas = ‘strong’ in Bontoc language]

Affixes are bound morphemes.

Prefixes

- No prefix can determine the category of a complex word:
- Eg: unhappy, unhappiness, unhappily, undo
- What does *un-* mean when it attaches to adjectives? un<u>un</u>kind, un<u>un</u>happy
- What does *un-* mean when it attaches to verbs?

un<u>un</u>do, un<u>un</u>tie

Suffixes

- We can represent the fact that the rightmost suffix determines the category of a word for triplets like -

Eg: rational, rationalize, rationalization

  - rational = adjective
  - rationalize = verb
  - rationalization = noun

Lexeme and Lemma

- In morphology, a lemma is the dictionary form of a set of words (headword)
- In English, for example, run, runs, ran and running are forms of the same lexeme, with **run as the lemma**.
- **Lexeme,** refers to the set of all the forms that have the same meaning, and
- **Lemma** refers to the particular form that is chosen by convention to represent the lexeme.

Stem and Lemma

- The **stem** is the part of the word that never changes even when morphologically inflected
- In linguistic analysis, the stem is defined as the analyzed base form from which all inflected forms can be formed.
- **For example**, from "produced", “production”, “product”, “products” **the stem is "produc-**".

- A **lemma** is the base form of the word.
- **For example,** from "produced", “production”, “product”, “products” **the lemma is "produce“**

Lemmatization

Lemmatization

- The process of determining the lemma for a given word is called **lemmatization**.
- The inflected forms of a word can be analyzed as a single item, identified by the word's **lemma**.
- For example, in English, the verb **'to walk'** may appear as 'walk', 'walked', 'walks', 'walking'.
- **The base form, 'walk', in a dictionary, is called the lemma for the word.**
- The association of the base form with a part of speech is called a **lexeme of the word**.

Lemmatization

- The word “**wake**" can be either the **base form of a noun or a form of a verb** ("to wake up") depending on the context.
- **Lemmatization** try to distinguish these two word senses, **stemming** would incorrectly conflate them.
- **Lemmatisation attempts to select the correct lemma depending on the context.**

The noun wake has 4 senses (first 1 from tagged texts)

1. (4) aftermath, wake, backwash -- (the consequences of an event (especially a catastrophic event); "the aftermath of war"; "in the wake of the accident no one knew how many had been injured")
2. Wake Island, Wake -- (an island in the western Pacific between Guam and Hawaii)
3. wake, backwash -- (the wave that spreads behind a boat as it moves forward; "the motorboat's wake capsized the canoe")
4. wake, viewing -- (a vigil held over a corpse the night before burial; "there's no weeping at an Irish wake")

The verb wake has 5 senses (first 2 from tagged texts)

1. (7) wake -- (be awake, be alert, be there)
2. (3) wake up, awake, arouse, awaken, wake, come alive, waken -- (stop sleeping; "She woke up to the sound of the alarm clock")
3. inflame, stir up, wake, ignite, heat, fire up -- (arouse or excite feelings and passions; "The ostentatious way of living of the rich ignites the hatred of the poor"; "The refugees' fate stirred up compassion around the world"; "Wake old feelings of hatred")
4. wake -- (make aware of; "His words woke us to terrible facts of the situation")
5. awaken, wake, waken, rouse, wake up, arouse -- (cause to become

Lemmatization

For instance:

1)The word "**better**" has "**good**" as its lemma. This link is missed by stemming, as it requires a dictionary look-up.

2)The word "**walk**" is the base form for word "**walking**", and hence this is matched in both stemming and lemmatisation.

Stemming

Stemming

- **Stemming** is the process of collapsing words into their morphological root.
- In linguistic analysis, the stem is defined as the analyzed base form from which all inflected forms can be formed. It is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form, usually root.
- **For example**, the terms addicted, addicting, addictions, addictive, and addicts might be conflated to their stem**, addict**.
- The process of stemming is often called conflation.

Example: Porter Stemmer, Snowball Stemmer, Wordnet Stemmer

Stemming Utility

The process of stemming is useful in search engines for Information retrieval

- Query expansion.
- Indexing.
- Natural language processing.

Differences between stemming and lemmatization

Differences between stemming and lemmatization

- Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.
- This indiscriminate cutting can be successful in some occasions, but not always, and that is why we affirm that this approach presents some limitations. Below is examples in English:

Form	Affixes	Lemma
Studies	-ies	Studi
Studying	-ing	Study

Differences between stemming and lemmatization

- Lemmatization, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

Form	Morphological Information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study

How do Stemming and Lemmatization work?

- **Stemming:** there are different algorithms that can be used in the stemming process, but the most common in English is **Porter stemmer**. The rules contained in this algorithm are divided in five different phases numbered from 1 to 5. The purpose of these rules is to reduce the words to the root.

- **Lemmatization:** the key to this methodology is linguistics. To extract the proper lemma, it is necessary to look at the morphological analysis of each word. This requires having dictionaries for every language to provide that kind of analysis.

Analyzing a word

- Look at the **word ACTORS**
- **Three morphemes**: act, -or, -s
- **Root:** act
- **Suffixes**: -or, -s
- **Derivational suffix**: -or
- **Inflectional suffix**: -s
- **Lexeme**: ACTOR
- **Word-forms**: actor, actors

Stemming Algorithms

Stemming Algorithms

- There are several types of stemming algorithms which differ in respect to performance and accuracy and how certain stemming obstacles are overcome.

- A stemmer for ENGLISH, for example, should identify the following :

1) STRING "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and
2) "stemmer", "stemming", "stemmed" as based on "stem".
3) "fishing", "fished", "fish", and "fisher" to the root word, "fish".

Brute Force Algorithms

- These stemmers employ a lookup table which contains relations between root forms and inflected forms. To stem a word, the table is queried to find a matching inflection. If a matching inflection is found, the associated root form is returned.

Advantages –

- Stemming error less.
- User friendly.

Problems –

- Time consuming.
- Back end updating
- Difficult to design.

Suffix Stripping Algorithms

- Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations.
- Instead, list of "rules" are maintained which provide a path for the algorithm, given an input word form to find its root form.
- Some examples of the rules include:
  - if the word ends in 'ed', remove the 'ed'
  - if the word ends in 'ing', remove the 'ing'
  - if the word ends in 'ly', remove the 'ly'

Lemmatization Algorithms

- Involves first determining the part of speech of a word, and then applying different normalization rules for each part of speech.
- The part of speech is first detected prior to attempting to find the root since for some languages, the rules change depending on a word's part of speech.

Hybrid Approaches

- Hybrid approaches use two or more of the approaches described above .
- A simple example is a algorithm which first consults a lookup table using brute force. However, instead of trying to store the entire set of relations between words in a given language, the lookup table is kept small and is only used to store a minute amount of "frequent exceptions" like "ran => run".
- If the word is not in the exception list, apply suffix stripping or lemmatization and output the result