Chapter NLP:III

III. Words

- Word-level Phenomena
- Text Representation
- Text Preprocessing
- Morphological Analysis
- Word Classes

Morphological Analysis

Overview [Hancox 1996]

Morphology is the study of the structure and formation of words.

call in the past $\xrightarrow{\text{inflection}}$ call ed

not being in a happy state $\xrightarrow{\text{derivation}}$ un happi ness

- A morpheme is a "minimal unit of meaning".
  Free morphemes can also be used as words.
  Bounded morphemes appear only as affixes (prefix, suffix, infix, and more) to words.

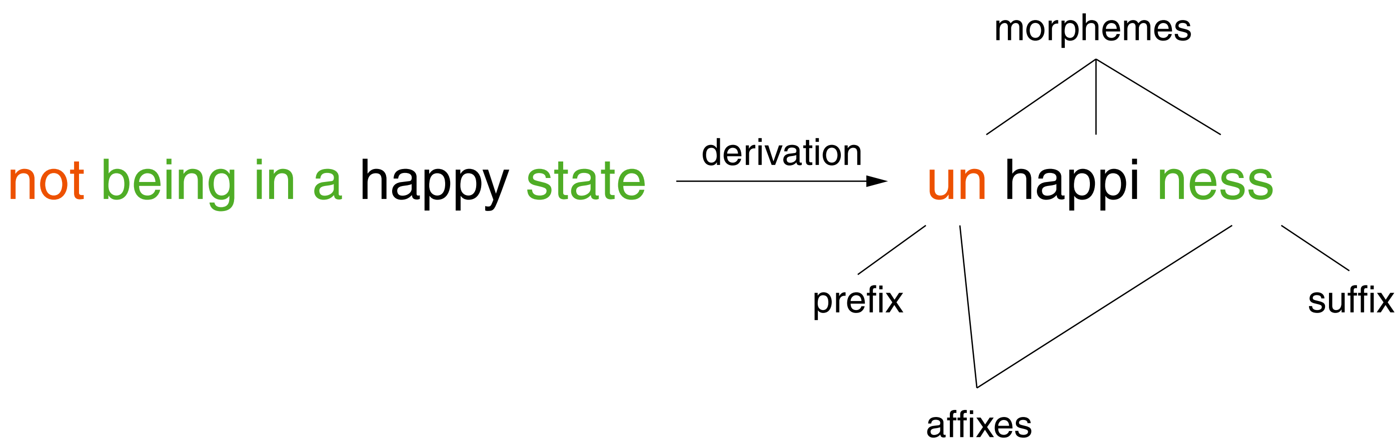
- A root is a single morpheme, a stem one or more.
  A root is the derivational base, or type, of a word, a stem its inflectional base.

→ Morphological analysis: identification of a word's morphemes and their role.

Morphological Analysis

Overview [Hancox 1996]

Morphology is the study of the structure and formation of words.



- A morpheme is a "minimal unit of meaning".
  Free morphemes can also be used as words.
  Bounded morphemes appear only as affixes (prefix, suffix, infix, and more) to words.

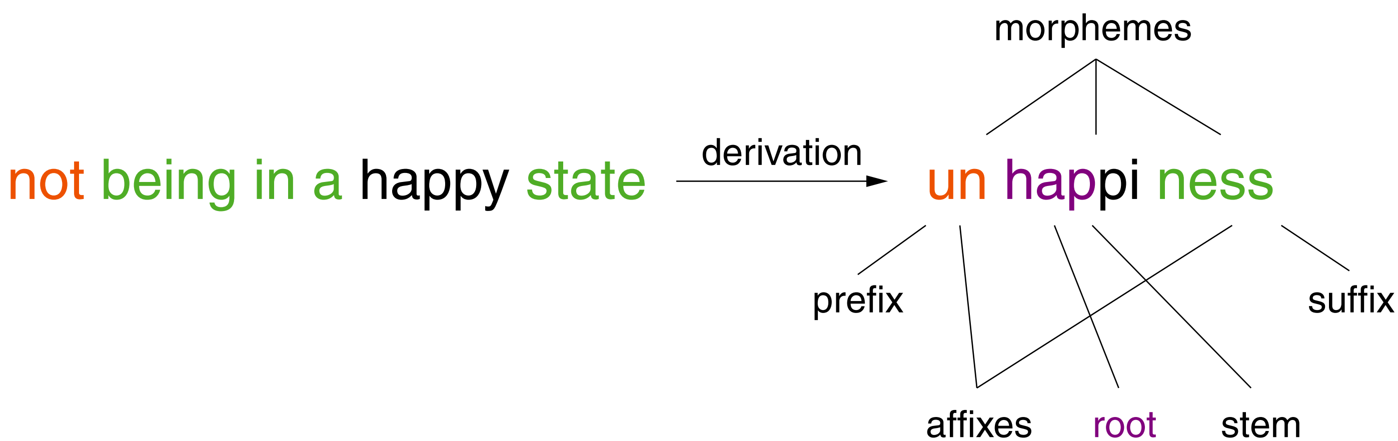
- A root is a single morpheme, a stem one or more.
  A root is the derivational base, or type, of a word, a stem its inflectional base.

→ Morphological analysis: identification of a word's morphemes and their role.

Morphological Analysis

Overview [Hancox 1996]

Morphology is the study of the structure and formation of words.



- A morpheme is a “minimal unit of meaning”.
  Free morphemes can also be used as words.
  Bounded morphemes appear only as affixes (prefix, suffix, infix, and more) to words.

- A root is a single morpheme, a stem one or more.
  A root is the derivational base, or type, of a word, a stem its inflectional base.

→ Morphological analysis: identification of a word’s morphemes and their role.

Morphological Analysis

Stemming

Mapping of a word token to its word stem by removal of inflection.

Inflections:

- noun declination (grammatical case, numerus, gender)
- verb conjugation (grammatical person, numerus, tense, mode, . . .)
- adjective and adverb comparison

A word stem forms the part of a word which remains after removing inflections (e.g., affixes).

Example:

```
connect    connects
           connected
           connecting
           connection
```


Morphological Analysis

Stemming: Principles [Frakes 1992]

1. Table lookup:
   Given a word stem, store its inflections in a hash table. Problem: completeness.

2. Affix elimination:
   Rule-based algorithms to identify prefixes and suffixes. Given their efficiency and intuitive workings, these are most commonly used.

3. Character n-grams:
   Usage of 4-grams or 5-grams from tokens as stems. Basic heuristic for English: use the first 4 characters as stem.

4. Successor variety:
   Exploits knowledge about structural linguistics to identify morpheme boundaries. The character sequences of tokens are added to a trie data structure; the outdegrees of inner nodes are analyzed to find suitable stems. Problem: difficult to operationalize.

Morphological Analysis

Stemming: Affix Elimination

Principle: “iterative longest match stemming”

1. Removal of the longest possible match based on a set of rules.
2. Repetition of Step 1 until no rule can be applied, anymore.
3. Recoding to address irregularities captured by the rules.

Morphological Analysis
Stemming: Affix Elimination

Principle: “iterative longest match stemming”

1. Removal of the longest possible match based on a set of rules.
2. Repetition of Step 1 until no rule can be applied, anymore.
3. Recoding to address irregularities captured by the rules.

Notation:

- $\mathtt{c}$ denotes a consonant, $\mathtt{C}$ a non-empty sequence of consonants.
  $\mathtt{v}$ denotes a vowel, $\mathtt{V}$ a non-empty sequence of vowels.
  → Every word is defined by $[\mathtt{C}](\mathtt{VC})^{\mathtt{m}}[\mathtt{V}]$
- Consonant: Letter that is not a vowel.
- Vowel: Letters A, E, I, O, and U as well as Y after a consonant.
  Example: In TOY the Y is a consonant, in LOVELY the Y is a vowel.

Morphological Analysis

Stemming: Porter Stemmer

Concepts:

- 9 rule sets, each consisting of 1-20 rules
- Rules of each group are sorted, to be applied top to bottom
- Only one rule per set can be applied
- Rules are defined as follows: `<Premise> S1` $\longrightarrow$ `S2`

Morphological Analysis

Stemming: Porter Stemmer

Concepts:

- 9 rule sets, each consisting of 1-20 rules
- Rules of each group are sorted, to be applied top to bottom
- Only one rule per set can be applied
- Rules are defined as follows: `<Premise> S1` $\longrightarrow$ `S2`

Semantics:

If a character sequence ends with `S1` and if the subsequence ahead of `S1` (= word stem) fulfills the `<Premise>`, replace `S1` by `S2`

Premises:

`(m>x)`	Number of vowel-consonant-sequences is larger than `x`.
`(*S)`	Word stem ends with `S`.
`(*v*)`	Word stem contains a vowel.
`(*o)`	Word stem ends with `cvc`, where the second consonant `c` $\notin$ `{W, X, Y}`.
`(*d)`	Word stem ends with two identical consonants.

Morphological Analysis

Stemming: Porter Stemmer

Selection of rules:

Rule set	Premise	Suffix	Replacement	Example
1a	Null	sses	ss	caresses $\rightarrow$ caress
1a	Null	ies	i	ponies $\rightarrow$ poni
1b	(m>0)	eed	ee	agreed $\rightarrow$ agree
				feed $\rightarrow$ feed
1b	(*v*)	ed	ε	plastered $\rightarrow$ plaster
				bled $\rightarrow$ bled
1b	(*v*)	ing	ε	motoring $\rightarrow$ motor
				sing $\rightarrow$ sing
1c	(*v*)	y	i	happy $\rightarrow$ happi
				sky $\rightarrow$ sky
2	(m>0)	biliti	ble	sensibiliti $\rightarrow$ sensible

Morphological Analysis

Stemming: Porter Stemmer

Example:

```
Alan Mathison Turing was an English mathematician, computer scientist,
logician, cryptanalyst, philosopher, and theoretical biologist.  Turing
was highly influential in the development of theoretical computer
science, providing a formalisation of the concepts of algorithm and
computation with the Turing machine, which can be considered a model of a
general-purpose computer.  Turing is widely considered to be the father
of theoretical computer science and artificial intelligence.
```

Morphological Analysis

Stemming: Porter Stemmer

Example:

Alan Mathison Tur**ing** wa**s** an English mathematician, comput**er** scientist, logician, cryptanalyst, philosoph**er**, and theoret**ical** biologist. Tur**ing** wa**s** highl**y** influenti**al** in the develop**ment** of theoret**ical** comput**er** scienc**e**, provid**ing** a formalis**ation** of the concepts of algorithm and computat**ion** with the Tur**ing** machin**e**, which can be consid**ered** a model of a gener**al**-purpos**e** comput**er**. Tur**ing** is wide**ly** consid**ered** to be the father of theoret**ical** comput**er** scienc**e** and artifici**al** intellig**ence**.

Morphological Analysis
Stemming: Porter Stemmer

Example:

`Alan Mathison Ture wa an English mathematician, comput scientist, logician, cryptanalyst, philosoph , and theoret biologist. Ture wa highli influenti in the develop of theoret comput scienc , provid a formalis of the concepts of algorithm and computat with the Ture machin , which can be consid a model of a gener -purpos comput . Ture is wide consid to be the father of theoret comput scienc and artifici intellig .`

Morphological Analysis

Stemming: Porter Stemmer

Weaknesses of the algorithm:

- Difficult to modify:
  The effects of changes are hardly predictable.

- Tends to overgeneralize:
  `university/universe, organization/organ`

- Does not capture clear generalizations:
  `European/Europe, matrices/matrix, machine/machineri`

- Generates word stems that are difficult to be interpreted:
  `iteration, general`

Morphological Analysis
Stemming: Krovetz Stemmer

The Krovetz stemmer combines a dictionary-based approach with rules. The dictionary captures well-known cases, whereas the rules capture words not known at the time of dictionary creation.

1. Word looked up in dictionary
2. If present, replaced with word stem
3. If not present, word is checked for removable inflection suffixes
4. After removal, dictionary is checked again
5. If still not present, different suffixes are tried

Observations:

- Captures irregular cases such as `is`, `be`, `was`.
- Produces words not stems (more readable, similar to lemmatization)
- Comparable effectiveness to Porter stemmer
- Lower false positive rate, somewhat higher false negative rate

Morphological Analysis

Stemming: Stemmer Comparison

Porter Stemmer:

```
Alan Mathison Turing was an English mathematician, computer scientist,
logician, cryptanalyst, philosopher, and theoretical biologist.  Turing
was highly influential in the development of theoretical computer
science, providing a formalisation of the concepts of algorithm and
computation with the Turing machine, which can be considered a model of a
general-purpose computer.  Turing is widely considered to be the father
of theoretical computer science and artificial intelligence.
```

Krovetz Stemmer:

```
Alan Mathison Turing was an English mathematician, computer scientist,
logician, cryptanalyst, philosopher, and theoretical biologist.  Turing
was highly influential in the development of theoretical computer
science, providing a formalisation of the concepts of algorithm and
computation with the Turing machine, which can be considered a model of a
general-purpose computer.  Turing is widely considered to be the father
of theoretical computer science and artificial intelligence.
```

Morphological Analysis

Stemming: Stemmer Comparison

Porter Stemmer:

Tur**ing** wa**s** an English mathematician, comput**er** scientist, logician, cryptanalyst, philosoph**er**, and theoret**ical** biologist. Tur**ing** wa**s** highl**y** influenti**al** in the develop**ment** of theoret**ical** comput**er** scienc**e**, provid**ing** a formalis**ation** of the concepts of algorithm and computat**ion** with the Tur**ing** machin**e**, which can be consid**ered** a model of a gener**al**-purpos**e** comput**er**. Tur**ing** is wide**ly** consid**ered** to be the father of theoret**ical** comput**er** scienc**e** and artifici**al** intellig**ence**.

Alan Mathison Tur**ing** was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist. Tur**ing** was highly influential in the development of theoretical computer science, providing a formalis**ation** of the concepts of algorithm and computation with the Tur**ing** machine, which can be considered a model of a general-purpose computer. Tur**ing** is widely considered to be the father of theoretical computer science and artificial intelligence.

Morphological Analysis

Stemming: Stemmer Comparison

Porter Stemmer:

Alan Mathison Tur**e** wa an English mathematician, comput scientist, logician, cryptanalyst, philosoph , and theoret biologist. Tur**e** wa highl**i** influenti in the develop of theoret comput scienc , provid a formalis of the concepts of algorithm and computat with the Tur**e** machin , which can be consid a model of a gener -purpos comput . Tur**e** is wide consid to be the father of theoret comput scienc and artifici intellig .

Krovetz Stemmer:

Alan Mathison Tur**e** was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist. Tur**e** was highly influential in the development of theoretical computer science, providing a formalis**e** of the concepts of algorithm and computation with the Tur**e** machine, which can be considered a model of a general-purpose computer. Tur**e** is widely considered to be the father of theoretical computer science and artificial intelligence.

Morphological Analysis

Stemming: Character n-grams [McNamee et al. 2004] [McNamee et al. 2008]

A substring of length n from a longer string is called a character n-gram. A string of length $m \geq n$ has at most $(m-n)+1$ distinct character n-grams.

Example: `Alan Mathison Turing ...`

- 1-grams: `A, l, a, n, M, a, t, h, i, s, o, n, T, u, r, i, n, g`
- 2-grams: `Al, la, an, Ma, at, th, hi, is, so, on, Tu, ur, ri, in, ng`
- 3-grams: `Ala, lan, Mat, ath, thi, his, iso, son, Tur, uri, rin, ing`
- 4-grams: `Alan, Math, athi, this, hiso, ison, Turi, urin, ring`
- 5-grams: `Alan, Mathi, athis, thiso, hison, Turin, uring`

Morphological Analysis

Stemming: Character n-grams [McNamee et al. 2004] [McNamee et al. 2008]

A substring of length n from a longer string is called a character n-gram. A string of length $m \geq n$ has at most $(m - n) + 1$ distinct character n-grams.

Example: `Alan Mathison Turing ...`

- 1-grams: `A, l, a, n, M, a, t, h, i, s, o, n, T, u, r, i, n, g`
- 2-grams: `Al, la, an, Ma, at, th, hi, is, so, on, Tu, ur, ri, in, ng`
- 3-grams: `Ala, lan, Mat, ath, thi, his, iso, son, Tur, uri, rin, ing`
- 4-grams: `Alan, Math, athi, this, hiso, ison, Turi, urin, ring`
- 5-grams: `Alan, Mathi, athis, thiso, hison, Turin, uring`

Use all character n-grams for $n = 4$ or $n = 5$ as pseudo-stems of a word.

Observations:

- Language-independent; good performance for many languages.
- Well-developed stemmers yield better performance (e.g., for English).
- Large overhead in terms of vocabulary size.