Morphology

Source:
Sudeshna Sarkar, IIT Kharagpur
Morphology

- **Morphology** is the field of linguistics that studies the internal structure of words.
- How words are built up from smaller meaningful units called morphemes (morph = shape, logos = word).
- We can usefully divide morphemes into two classes:
  - **Stems**: The core meaning bearing units.
  - **Affixes**: Bits and pieces that adhere to stems to change their meanings and grammatical functions.
    - Prefix: un-, anti-, etc (a- ati- pra- etc)
    - Suffix: -ity, -ation, etc (-taa, -ke, -ka etc)
    - Infix: are inserted inside the stem
      - Tagalog: um + hingi → humingi
    - Circumfixes – precede and follow the stem.
- Turkish can have words with a lot of suffixes (agglutinative language).
- Many Indian languages also have agglutinative suffixes.
Morpheme Definitions

- **Root**
  - The portion of the word that:
    - is common to a set of derived or inflected forms, if any, when all affixes are removed
    - is not further analyzable into meaningful elements
    - carries the principle portion of meaning of the words

- **Stem**
  - The root or roots of a word, together with any derivational affixes, to which inflectional affixes are added.

- **Affix**
  - A bound morpheme that is joined before, after, or within a root or stem.

- **Clitic**
  - A morpheme that functions syntactically like a word, but does not appear as an independent phonological word
    - Spanish: *un beso, las aguas*, English: *Hal’s* (genetive marker)
Inflectional Morphology

• **Inflection:**
  – Variation in the form of a word, typically by means of an affix, that expresses a grammatical contrast.
    • Doesn’t change the word class
    • Usually produces a predictable, nonidiosyncratic change of meaning.
    • Serves a grammatical/semantic purpose different from the original

After a combination with an **inflectional morpheme**, the meaning and class of the actual stem usually do not change.
  – eat / eats , pencil / pencils
Inflectional Morphology

• Adds:
  – tense, number, person, mood, aspect
• Word class doesn’t change
• Word serves new grammatical role
• Examples
  – *come* is inflected for person and number:
    \[ \text{The pizza guy comes at noon.} \]
  – *las* and *rojas* are inflected for agreement with *manzanas* in grammatical gender by -*a* and in number by –*s
    \[ \text{las manzanas rojas} \quad \text{(`the red apples' \text{)})} \]
• Derivation:
  – The formation of a new word or inflectable stem from another word or stem.

• After a combination with a derivational morpheme, the meaning and the class of the actual stem usually change.
  – compute / computer  do / undo  friend / friendly
  – Uygar / uygarlaş  kapı / kapıçı
  – udaara (J) / udaarataa (N)
  – bhadra / abhadra
  – baayu / baayabiiya

• Irregular changes may happen with derivational affixes.
Derivational Morphology

• Nominalization (formation of nouns from other parts of speech, primarily verbs in English):
  – computerization
  – appointee
  – killer
  – fuzziness

• Formation of adjectives (primarily from nouns)
  – computational
  – clueless
  – Embraceable
Concatinative Morphology

- Morpheme+Morpheme+Morpheme+…
- Stems: also called lemma, base form, root, lexeme
  - hope+ing ➔ hoping  hop ➔ hopping
- Affixes
  - Prefixes: Antidisestablishmentarianism
  - Suffixes: Antidisestablishmentarianism
  - Infixes: hingi (borrow) – humingi (borrower) in Tagalog
  - Circumfixes: sagen (say) – gesagt (said) in German
- Agglutinative Languages
  - uygarlaştıramadıklarımızdanmışsınızcasına
  - uygar+la+ş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına
  - Behaving as if you are among those whom we could not cause to become civilized
Templatic Morphology

- **Roots and Patterns**
  - **Example:** Hebrew verbs
  - **Root:**
    - Consists of 3 consonants CCC
    - Carries basic meaning
  - **Template:**
    - Gives the ordering of consonants and vowels
    - Specifies semantic information about the verb
      - Active, passive, middle voice
  - **Example:**
    - lmd (to learn or study)
      - CaCaC -> lamad (he studied)
      - CiCeC -> limed (he taught)
      - CuCaC -> lumad (he was taught)
Syntax and Morphology

• Phrase-level agreement
  – Subject-Verb
    • John studies hard \((\text{STUDY}+\text{3SG})\)
  – Noun-Adjective
    • Las vacas hermosas

• Sub-word phrasal structures
  – That+in+book+PL+Poss:1PL
  – Which are in our books
Surface and Lexical Forms

- The **surface level** of a word represents the actual spelling of that word.
  - geliyorum eats cats kitabım
- The **lexical level** of a word represents a simple concatenation of morphemes making up that word.
  - gel +PROG +1SG
  - eat +AOR
  - cat +PLU
  - kitap +P1SG
- Morphological processors try to find correspondences between lexical and surface forms of words.
  - **Morphological recognition/** analysis – surface to lexical
  - **Morphological generation/** synthesis – lexical to surface
Morphology: Morphemes & Order

• Handles what is an isolated form in written text

• Grouping of phonemes into morphemes
  – sequence deliverables ~ deliver, able and s (3 units)

• Morpheme Combination
  – certain combinations/sequencing possible, other not:
    • deliver+able+s, but not able+derive+s; noun+s, but not noun+ing
    • typically fixed (in any given language)
Morphological Parsing

• Morphological parsing is to find the lexical form of a word from its surface form.
  - cats -- cat +N +PLU
  - cat -- cat +N +SG
  - goose -- goose +N +SG or goose +V
  - geese -- goose +N +PLU
  - gooses -- goose +V +3SG
  - catch -- catch +V
  - caught -- catch +V +PAST or catch +V +PP
  - AsachhilAma AsA+PROG+PAST+1st I/We was/were coming

• There can be more than one lexical level representation for a given word. (ambiguity)
  flies
    fly\text{VERB}+PROG
    fly\text{NOUN}+PLU
  mAtAla
  kare
Formal definition of the problem

- **Surface form**: The word \( w_s \) as it occurs in the text. [sings]
  \[
  w_s \in L \subseteq \Sigma^+
  \]

- **Lexical form**: The root word(s) \( (r_1, r_2, \ldots) \) and other grammatical features \( (F) \). [sing,v,+sg,+3rd]
  \[
  w_l \in \{\Sigma^+,\}^+F^+
  \]

  \[
  w_l \in \Delta^+
  \]
Analysis & Synthesis

- **Morphological Analysis**: Maps a string from surface form to corresponding lexical form.
  \[ f_{MA}: \Sigma^+ \rightarrow \Delta^+ \]

- **Morphological Synthesis**: Maps a string from lexical form to surface form.
  \[ f_{MA}: \Delta^+ \rightarrow \Sigma^+ \]
Relationship between MA & MS

\[ f_{MS} \circ f_{MA}(w_s) = w_s \]
\[ f_{MA} \circ f_{MS}(w_l) = w_l \]

\[ f_{MS} = f_{MA}, \quad f_{MA} = f_{MS} \]

But is that really the case?
Example

• Fly + s → flys → flies (y → i rule)
• Duckling

Go-getter → get + er
Doer → do + er
Beer → ?

What knowledge do we need?
How do we represent it?
How do we compute with it?
Knowledge needed

• Knowledge of stems or roots
  – Duck is a possible root, not duckl
  We need a dictionary (lexicon)
• Only some endings go on some words
  – Do + er ok
  – Be + er – not ok
• In addition, spelling change rules that adjust the surface form
  – Get + er – double the t getter
  – Fox + s – insert e – foxes
  – Fly + s – insert e – flys – y to i – flies
  – Chase + ed – drop e - chased
Put all this in a big dictionary (lexicon)

- Turkish – approx $600 \times 10^6$ forms
- Finnish – $10^7$
- Hindi, Bengali, Telugu, Tamil?
- Besides, always novel forms can be constructed
  - Anti-missile
    - Anti-anti-missile
      - Anti-anti-anti-missile
        » ........
- Compounding of words – Sanskrit, German
Morphology: From Morphemes to Lemmas & Categories

• **Lemma**: lexical unit, “pointer” to lexicon
  – typically is represented as the “base form”, or “dictionary headword”
  • possibly indexed when ambiguous/polysemous:
    – state\(^1\) (verb), state\(^2\) (state-of-the-art), state\(^3\) (government)
  – from one or more morphemes (“root”, “stem”, “root+derivation”, …)

• **Categories**: non-lexical
  – small number of possible values (< 100, often < 5-10)
Morphology Level: The Mapping

- Formally: \( A^+ \rightarrow 2^{(L,C_1,C_2,...,C_n)} \)
  - \( A \) is the alphabet of phonemes (\( A^+ \) denotes any non-empty sequence of phonemes)
  - \( L \) is the set of possible lemmas, uniquely identified
  - \( C_i \) are morphological categories, such as:
    - grammatical number, gender, case
    - person, tense, negation, degree of comparison, voice, aspect, ...
    - tone, politeness, ...
    - part of speech (not quite morphological category, but...)

- \( A, L \) and \( C_i \) are obviously language-dependent
Morphological Analysis (cont.)

- Relatively simple for English.
- But for many Indian languages, it may be more difficult.

Examples

Inflectional and Derivational Morphology.

- Common tools: Finite-state transducers
  - A transducer maps a set/string of symbols to another set/string of symbols
A simpler problem

• Linear concatenation of morphemes with possible spelling changes at the boundary and a few irregular cases.

• Quite practical assumptions
  – English, Hindi, Bengali, Telugu, Tamil, French, Turkish …
  – Exceptions: Semitic languages, Sanskrit
Computational Morphology

• Approaches
  – Lexicon only
  – Rules only
  – Lexicon and Rules
    • Finite-state Automata
    • Finite-state Transducers
Computational Morphology

• Systems
  – WordNet’s morphy
  – PCKimmo
    • Named after Kimmo Koskenniemi, much work done by Lauri Karttunen, Ron Kaplan, and Martin Kay
    • Accurate but complex
    • [http://www.sil.org/pckimmo/](http://www.sil.org/pckimmo/)
  – Two-level morphology
    • Commercial version available from InXight Corp.

• Background
  – Chapter 3 of Jurafsky and Martin
  – A short history of Two-Level Morphology
Finite State Machines

- FSAs are equivalent to regular languages
- FSTs are equivalent to regular relations (over pairs of regular languages)
- FSTs are like FSAs but with complex labels.
- We can use FSTs to *transduce* between surface and lexical levels.
Can FSAs help?

Q₀ → Reg-noun → Q₁ → Plural (-s) → Q₂

Irreg-pl-noun → Irreg-sg-noun
What's this for?

$un \, \text{Adj-root} \{\text{er} / \text{est} / \text{ly}\}$?
Morphotactics

• The last two examples basically model some parts of the English *morphotactics*

• But where is the information about regular and irregular roots?

LEXICON

• Can we include the lexicon in the FSA?
The English Pluralization FSA
After adding a mini-lexicon
Finite State Transducers

Surface form

Lexical form

Finite State Machine

<table>
<thead>
<tr>
<th>s</th>
<th>i</th>
<th>n</th>
<th>g</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>i</td>
<td>n</td>
<td>g</td>
<td>#</td>
</tr>
</tbody>
</table>
Formal Definition

- A 6-tuple \( \{ \Sigma, \Delta, Q, \delta, q_0, F \} \)
  - \( \Sigma \) is the (finite) set of input symbols
  - \( \Delta \) is the (finite) set of output symbols
  - \( Q \) is the set (FINITE) of states
  - \( \delta \) is the transition function \( Q \times \Sigma \) to \( Q \times \Delta \)
  - \( q_0 \in Q \) is the start state
  - \( F \subseteq Q \) is the set of accepting states
An example FST
The Lexicon FST

Graphical representation of a Finite State Transducer (FST)
Modelling Orthographic Rules

• Spelling changes in morpheme boundaries
  – bus+s $\rightarrow$ busses, watch+s $\rightarrow$ watches
  – fly+s $\rightarrow$ flies
  – make+ing $\rightarrow$ making

• Rules
  – E-insertion takes place if the stem ends in s, z, ch, sh etc.
  – y maps to ie when pluralization marker s is added
Rewrite Rules

• Chomsky and Halle (1968)
• General form:

\[ a \rightarrow b / \lambda \_\_ \rho \]

• E-insertion:

\[ \varepsilon \rightarrow e / \{x,s,z,ch,sh\ldots\}^\_\_ s# \]

• Kay and Kaplan (1994) showed that FSTs can be compiled from general rewrite rules
Two-level Morphology (Koskenniemi, 1983)

b | u | s | +N | +Pl

LEXICON FST

b | u | s | ^ | s | #

FST₁ orthographic rules FSTₙ

b | u | s | e | s
A Single FST for MA and MS

LEXICON FST

FST$_1$

orthographic rules

FST$_n$

Morphology FST
Can we do without the lexicon

• Not really!
• But for some applications we might need to know the stem only
• Surface form $\rightarrow$ Stem [Stemming]
• Porter Stemming algorithm (1980) is a very popular technique that does not use lexicon.
Other Issues

• How to formulate the rewrite rules?
• How to ensure coverage?
• What to do for unknown roots?
• Is it possible to learn morphology of a language in supervised/unsupervised manner?
• What about non-linear morphology?
Derivational Rules
Morphological Analyzer

To build a morphological analyser we need:

• **lexicon**: the list of stems and affixes, together with basic information about them

• **morphotactics**: the model of morpheme ordering (e.g., English plural morpheme follows the noun rather than a verb)

• **orthographic rules**: these spelling rules are used to model the changes that occur in a word, usually when two morphemes combine (e.g., fly+s = flies)
Lexicon & Morphotactics

• Typically list of word parts (lexicon) and the models of ordering can be combined together into an FSA which will recognise all the valid word forms.
• For this to be possible the word parts must first be classified into sublexicons.
• The FSA defines the morphotactics (ordering constraints).
Towards the Analyser

• We can use lexc or xfst to build such an FSA (see lex1.lexc)

• To augment this to produce an analysis we must create a transducer $T_{num}$ which maps between the lexical level and an "intermediate" level that is needed to handle the spelling rules of English.
Three Levels of Analysis

**Lexical**

```
fox +N +PL
```

**Intermediate**

```
fox ^s #
```

**Surface**

```
foxes
```
1. $T_{num}$: Noun Number Inflection

- multi-character symbols
- morpheme boundary ^
- word boundary #
The reason we need to have an intermediate form is that funny things happen at morpheme boundaries, e.g.

\[\text{cat}^s \leftrightarrow \text{cats}\]
\[\text{fox}^s \leftrightarrow \text{foxes}\]
\[\text{fly}^s \leftrightarrow \text{flies}\]

The rules which describe these changes are called orthographic rules or "spelling rules".
More English Spelling Rules

- **consonant doubling**: beg / begging
- **y replacement**: try/tries
- **k insertion**: panic/panicked
- **e deletion**: make/making
- **e insertion**: watch/watches

Each rule can be stated in more detail ...
Spelling Rules

• Chomsky & Halle (1968) invented a special notation for spelling rules.

• A very similar notation is embodied in the "conditional replacement" rules of xfst.

\[ E \rightarrow F \mid \mid L \_ R \]

which means replace E with F when it appears between left context L and right context R
This rule does e-insertion

\^ \rightarrow e \mid \mid x \_ s#
e insertion over 3 levels

The rule corresponds to the mapping between surface and intermediate levels

**Lexical**

```
\{ f o x +N +PL \}
```

**Intermediate**

```
\{ f o x ^s # \}
```

**Surface**

```
\{ f o x e s \}
```
e insertion as an FST
Incorporating Spelling Rules

• Spelling rules, each corresponding to an FST, can be run *in parallel* provided that they are "aligned".
• The set of spelling rules is positioned between the surface level and the intermediate level.
• Parallel execution of FSTs can be carried out:
  – by simulation: in this case FSTs must first be aligned.
  – by first constructing a single FST corresponding to their intersection.
Adding in the Words
Derivational Rules

Diagram showing derivational rules with transitions labeled by language features such as noun, verb, adjective, and suffixes like -ive, -ful, -ly, -ness, -able, -ity, -er, -izer, -ization, and -ation.
Parsing/Generation vs. Recognition

• Recognition is usually not quite what we need.
  – Usually if we find some string in the language we need to find the structure in it (parsing)
  – Or we have some structure and we want to produce a surface form (production/generation)

• Example
  – From “cats” to “cat +N +PL” and back
Morphological Parsing

- Given the input *cats*, we’d like to output
  *cat +N +Pl*,
telling us that *cat* is a plural noun.
- Given the Spanish input *bebo*, we’d like to output
  *
  *beber +V +PlInd +1P +Sg*
telling us that *bebo* is the present indicative first person singular form of the Spanish verb
  *beber*, ‘to drink’.
Lexicon & Morphotactics

• Typically list of word parts (lexicon) and the models of ordering can be combined together into an FSA which will recognise all the valid word forms.

• For this to be possible the word parts must first be classified into sublexicons.

• The FSA defines the morphotactics (ordering constraints).
Putting it all together

Lexical  LEXICON-FST

Intermediate

Surface

execution of FST$_i$ takes place in parallel
Ambiguity

- Recall that in non-deterministic recognition multiple paths through a machine may lead to an accept state.
  - Didn’t matter which path was actually traversed

- In FSTs the path to an accept state does matter since differ paths represent different parses and different outputs will result
Ambiguity

• What’s the right parse for
  – Unionizable
  – Union-ize-able
  – Un-ion-ize-able

• Each represents a valid path through the derivational morphology machine.
• There are a number of ways to deal with this problem
  – Simply take the first output found
  – Find all the possible outputs (all paths) and return them all (without choosing)
  – Bias the search so that only one or a few likely paths are explored
The Gory Details

• Of course, it's not as easy as
  − “cat +N +PL” <-> “cats”

• As we saw earlier there are geese, mice and oxen

• But there are also a whole host of spelling/pronunciation changes that go along with inflectional changes
  − Cats vs Dogs
  − Fox and Foxes
Multi-Level Tape Machines

- We use one machine to transduce between the lexical and the intermediate level, and another to handle the spelling changes to the surface tape.
Lexical to Intermediate Level
Intermediate to Surface

- The add an “e” rule as in $\text{fox}^*\text{s}\# \leftrightarrow \text{foxes}\#$
Foxes
A key feature of this machine is that it doesn’t do anything to inputs to which it doesn’t apply.

Meaning that they are written out unchanged to the output tape.

Turns out the multiple tapes aren’t really needed; they can be compiled away.
Overall Scheme

• We now have one FST that has explicit information about the lexicon (actual words, their spelling, facts about word classes and regularity).
  – Lexical level to intermediate forms

• We have a larger set of machines that capture orthographic/spelling rules.
  – Intermediate forms to surface forms
Overall Scheme

Lexical

Intermediate

Surface

LEXICON-FST

FST$_1$

orthographic rules

FST$_n$