Computational Linguistics 1
CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland
Lecture 3: 8 September 2011

Agenda
- HW1 – online tonight, due next Thursday
- Morphology
  - Words
  - Concatenative vs. non-concatenative
  - Inflectional vs. derivational morphology
  - Regular vs. irregular
  - Formal morphology
- Computational morphology
  - Finite-state methods

Morphology
- Study of how words are constructed from smaller units of meaning
- Smallest unit of meaning = morpheme
  - fox has morpheme fox
  - cats has two morphemes cat and -s
- Note: it is useful to distinguish morphemes from orthographic rules
- Two classes of morphemes:
  - Stems: supply the “main” meaning
  - Affixes: add “additional” meaning

Why Morphology?
- In English, morphology is relatively impoverished
- Nevertheless, even in English there are some important uses, e.g.,
  - Features for processing OOV words
  - Stemming for document classification
- Usually very simple techniques suffice in English (e.g., Porter stemmer: may be wrong, but systematically wrong)
- Very accurate non-statistical algorithms exist for English
- Other languages, such as Turkish, require a more serious morphological processing

Issues in Morphology
- What is a word?
- What kinds of things can words encode?
- How are words put together?

Words
- Orthographic word:
  Words as defined by delimiters in written text.
- Sociological word:
  “The unit, intermediate in size between a phoneme and a sentence, which the general, non-linguistic public is conscious of and has an everyday term for.”
- Morphological word:
  Anything that is the output of a word-formation rule.
- Lexical, semantic, phonological, syntactic, psycholinguistic definitions of “word”
Meaning Encoded by Morphology

Mohawk (Baker, 1996):
- Ra-wis-subwe’s
- MoS-baby-B-like-HAB
  *He likes babies*

Alaskan Yupik (Woodbury 1987):
- qeqliiyiitik
  *be excellent (-y-) at making (-k-) kayaks (qay-)*

Turkish (Hankamer, 1986):
- önpükünürünüdenkennedi
  *(garbage=AFF+PL+IP.PL=LOC+REL+PL=ABL+INT+AUX+PAST)*
  *was it from those that were in our garbage cans?*

Topology of Morphologies

- Concatenative vs. non-concatenative
- Derivational vs. inflectional
- Regular vs. irregular

Inflection vs. Derivation vs. Compounding

- Concatenative forms new words by adding to a stem word
- Inflection yields new forms of the same word
  - tense, number, mood, voice marking in verbs
  - case, number, gender marking in nominals
  - comparison of adjectives (e.g., big bigger biggest)
- Derivation yields different words
  - Derived nominals
  - Derived adjectives
  - Derived verbs
  - (adjectives & verbs derived from nouns)
- Compounding forms new words out of 2+ other words
  - Noun-noun compounding
  - Incorporation

Concatenative Morphology

- Morpheme+Morpheme+Morpheme+…
- Stems (also called lemma, base form, root, lexeme):
  - hope+ing → hoping
  - hopping
- Affixes:
  - Prefixes: Anti-disestablishmentarianism
  - Suffixes: Anti-disestablishmentarianism
- Agglutinative languages (e.g., Turkish)
  - uyardırıtmadıklarımızdan mı şınızcasına → uyardırı+ma+da+mı+ss+an+ca
  - Meaning: behaving as if you are among those whom we could not cause to become civilized

Non-Concatenative Morphology

- Infixes (e.g., Tagalog)
  - hingi (borrow)
  - humingi (borrower)
- Circumfixes (e.g., German)
  - sagen (say)
  - gesagt (said)
- Reduplication (e.g., Motu, spoken in Papua New Guinea)
  - mahuta (to sleep)
  - mahutahuta (to sleep constantly)
  - mahutahuta (to sleep, plural)

Infixed Examples

- (vs pre-fixation/suffixation of concatenative morphology)
  - Bonnice (Pembroke and Rodman 1983): Do normal affixation, ignoring a segment:
    - skas: strong
    - kudal: hot
    - fusial: enemy
  - Bongo (CODIUL 1989): Perse out a prosodic unit here a foot-and attach to it:
    - bilan bilamir ‘fish’/‘my fish’
    - dili dihikmin ‘snake’/‘my snake’
    - lima liklima ‘lense’/‘my lense’
    - slikhil slikhil ‘housy’/‘my housy’

English: attach after a foot:
- absolutely also -sually
- Kalamazoo -zo
Circumfixation Examples

säuseln ‘rustle’ gesäuselt ‘rustled’
brüsten ‘brag’ gebrüstet ‘bragged’
täuschen ‘deceive’ getäuscht ‘deceived’

• Note: circumfixation involves long distance dependency: the -t needs to remember that a ge- has been seen.

Templatic Morphologies

• Common in Semitic languages
• Roots and patterns

<table>
<thead>
<tr>
<th>Arabic</th>
<th></th>
<th>Hebrew</th>
</tr>
</thead>
<tbody>
<tr>
<td>كتب</td>
<td>مكتوب</td>
<td>کتب</td>
</tr>
<tr>
<td>كتب</td>
<td>written</td>
<td>كتب</td>
</tr>
</tbody>
</table>

More Templatic Morphology

<table>
<thead>
<tr>
<th>Binyan active</th>
<th>passive</th>
<th>template</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>katab</td>
<td>kutib</td>
<td>CVCVC</td>
</tr>
<tr>
<td>II</td>
<td>katab</td>
<td>kuttib</td>
<td>CVCCVC</td>
</tr>
<tr>
<td>III</td>
<td>kaatab</td>
<td>kuutib</td>
<td>CVVCVC</td>
</tr>
<tr>
<td>VI</td>
<td>taakaatab</td>
<td>tuukuutib</td>
<td>nCVVCVC</td>
</tr>
<tr>
<td>VII</td>
<td>nkaaab</td>
<td>nkkuuib</td>
<td>nCVVCVC</td>
</tr>
<tr>
<td>VIII</td>
<td>ktaaab</td>
<td>kuutib</td>
<td>CVCVC</td>
</tr>
<tr>
<td>X</td>
<td>staktab</td>
<td>stuktiib</td>
<td>stVCVCVC</td>
</tr>
</tbody>
</table>

Inflectional Morphology

• Stem + morpheme →
  • Word with same part of speech as the stem
  • Adds: tense, number, person,…
• Plural morpheme for English noun
  • cat+s
  • dog+s
• Progressive form in English verbs
  • walk+ing
  • rain+ing

Inflectional Categories

• Most languages mark case
  • if not morphologically, by syntactic means (e.g., prepositions)
• Many languages lack morphological gender
• Many languages lack systematic marking for number
• Many languages that have some of these markings…
  • still lack agreement for these features
• In many cases (as in Latin) some forms serve multiple functions
  • Ambiguity!

Noun Inflections in English

• Regular
  • cat/cats
  • dog/dogs
• Irregular
  • mouse/mice
  • ox/oxen
  • goose/geese
Verb Inflections in Spanish

<table>
<thead>
<tr>
<th>Present Indicative</th>
<th>Imperfect Indicative</th>
<th>Future</th>
<th>Imperfect Subjunctive</th>
<th>Present Subjunctive</th>
<th>Conditional Subjunctive</th>
<th>Future Subjunctive</th>
</tr>
</thead>
<tbody>
<tr>
<td>tengo</td>
<td>teniendo</td>
<td>tendré</td>
<td>tengamos</td>
<td>tengan</td>
<td>tendremos</td>
<td>tendrán</td>
</tr>
<tr>
<td>ayudas</td>
<td>ayudando</td>
<td>ayudará</td>
<td>ayudemos</td>
<td>ayuden</td>
<td>ayudaremos</td>
<td>ayudarán</td>
</tr>
<tr>
<td>estudia</td>
<td>estudiando</td>
<td>estudiar</td>
<td>estudiamos</td>
<td>estudien</td>
<td>estudiemos</td>
<td>estudiarán</td>
</tr>
<tr>
<td>aprende</td>
<td>aprendiendo</td>
<td>aprender</td>
<td>aprendamos</td>
<td>aprendan</td>
<td>aprendamos</td>
<td>aprendan</td>
</tr>
<tr>
<td>aprendemos</td>
<td>aprendemos</td>
<td>aprendan</td>
<td>aprendan</td>
<td>aprendan</td>
<td>aprendan</td>
<td>aprendan</td>
</tr>
<tr>
<td>aprended</td>
<td>aprended</td>
<td>aprended</td>
<td>aprended</td>
<td>aprended</td>
<td>aprended</td>
<td>aprended</td>
</tr>
<tr>
<td>aprendedlos</td>
<td>aprendedlos</td>
<td>aprended</td>
<td>aprendedlos</td>
<td>aprendedlos</td>
<td>aprendedlos</td>
<td>aprendedlos</td>
</tr>
<tr>
<td>aprendedlas</td>
<td>aprendedlas</td>
<td>aprended</td>
<td>aprendedlas</td>
<td>aprendedlas</td>
<td>aprendedlas</td>
<td>aprendedlas</td>
</tr>
</tbody>
</table>

Examples of Derivational Morphology

- Agentive Nominals
  - adder, banker, catcher, dealer, eater, fighter, grinder, hunter, ionizer, jumper, killer, lover, manager, name, opener, quitter . . .
  - Note: this function is marked using separate words in some languages. Cf. Mandarin zhe as in chi xigua zhe (eat watermelon AGENTIVE) ‘the one who is eating watermelon’
- Derived nominals
  - The Romans’ destruction of Carthage
  - In Mandarin there are no markings for this: verb phrases can simply function as nominals.
- Deadjectival nominals
  - rare/rarity, grammatical/grammaticality, grave/gravity
- Compound-like prefixes
  - pseudo-leftist, pseudoscience, pseudointellectual; semi-arid, semi-divine, semiregular

Compound Morphology

- firefighter, football, firecracker, policeman, doghouse
- Lebensversicherungsgeellschaftsangestellter ‘life insurance company employee’
- computer communications network performance analysis primer

Formal Morphology

- How is information formally encoded morphologically?
  - prefixation, suffixation
  - infixed
  - circumcision
  - templatic morphology
  - reduplication
  - subsegmental morphology
  - zero’ morphology
- What do these mean from a computational point of view?
Agenda

• HW1 – online tonight, due next Thursday
• Morphology
  • Words
  • Concatenative vs. non-concatenative
  • Inflectional vs. derivational morphology
  • Regular vs. irregular
  • Formal morphology
  • Computational morphology
    • Finite-state methods

Morphological Parsing

• Computationally decompose input forms into component morphemes
• Components needed:
  • A lexicon (stems and affixes)
  • A model of how stems and affixes combine
  • Orthographic rules

Morphological Parsing: Examples

<table>
<thead>
<tr>
<th>WORD</th>
<th>STEM (+FEATURES)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>cats</td>
<td>cat +N +PL</td>
</tr>
<tr>
<td>cat</td>
<td>cat +N +SG</td>
</tr>
<tr>
<td>cities</td>
<td>city +N +PL</td>
</tr>
<tr>
<td>geese</td>
<td>goose +N +PL</td>
</tr>
<tr>
<td>ducks</td>
<td>(duck +N +PL) or (duck +V +3SG)</td>
</tr>
<tr>
<td>merging</td>
<td>merge +V +PRES-PART</td>
</tr>
<tr>
<td>caught</td>
<td>(catch +V +PAST-PART) or (catch +V +PAST)</td>
</tr>
</tbody>
</table>

Different Approaches

• Lexicon only
• Rules only
• Lexicon and rules
  • finite-state automata
  • finite-state transducers

Lexicon-only

• Simply enumerate all surface forms and analyses
• So what’s the problem?
• When might this be useful?

acclaim acclaim $N$
acclaim acclaim $V$+0
acclaimed acclaim $V$+ed
acclaimed acclaim $V$+en
acclaiming acclaim $V$+ing$+$
acclaims acclaim $N$+s
acclaims acclaim $V$+s
acclamations acclaimation $N$+s
acclamations acclaimation $N$+ed
acclimated acclaimate $V$+0
acclimated acclaimate $V$+ed
acclimated acclaimate $V$+en$+$
acclimating acclaimate $V$+ing$+$

Rule-only: Porter Stemmer

• Cascading set of rules
  • ational $\rightarrow$ ate (e.g., relational $\rightarrow$ relate)
  • ing $\rightarrow$ e (e.g., walking $\rightarrow$ walk)
  • sses $\rightarrow$ ss (e.g., grasses $\rightarrow$ grass)
  • ...
• Examples
  • cities $\rightarrow$ citi
  • city $\rightarrow$ citi
  • generalizations
    • $\rightarrow$ generalization
    • $\rightarrow$ generalize
    • $\rightarrow$ general
    • $\rightarrow$ gener
Porter Stemmer: What's the Problem?

- Errors...

<table>
<thead>
<tr>
<th>Errors of Commission</th>
<th>Errors of Omission</th>
</tr>
</thead>
<tbody>
<tr>
<td>organization organ</td>
<td>European Europe</td>
</tr>
<tr>
<td>doing doe</td>
<td>analysis analyzes</td>
</tr>
<tr>
<td>numerical numerous</td>
<td>noise noising</td>
</tr>
<tr>
<td>policy police</td>
<td>sparse sparsity</td>
</tr>
</tbody>
</table>

- Why is it still useful?

Lexicon + Rules

- FSA: for recognition
  - Recognize all grammatical input and only grammatical input
- FST: for analysis
  - If grammatical, analyze surface form into component morphemes
  - Otherwise, declare input ungrammatical

FSA: English Noun Morphology

reg-noun | irreg-pl-noun | irreg-sg-noun | plural
---|---|---|---
fox | geese | goose | -s
cat | sheep | sheep | -
dog | mice | mouse | -

Note problem with orthography!

FSA: English Verb Morphology

<table>
<thead>
<tr>
<th>reg-verb-stem</th>
<th>irreg-verb-stem</th>
<th>irreg-past-verb</th>
<th>past</th>
<th>past-part</th>
<th>pres-part</th>
<th>3sg</th>
</tr>
</thead>
<tbody>
<tr>
<td>walk</td>
<td>fry</td>
<td>talk</td>
<td>impeach</td>
<td>cut</td>
<td>speak</td>
<td>spoken</td>
</tr>
</tbody>
</table>

FSA: English Adjectival Morphology

- Examples:
  - big, bigger, biggest
  - smaller, smaller, smallest
  - happy, happier, happiest, happily
  - unhappy, unhappier, unhappiest, unhappily
- Morphemes:
  - Roots: big, small, happy, etc.
  - Affixes: un-, -er, -est, -ly
**FSA: English Adjectival Morphology**

- $q_0$, $q_1$, $q_2$, $q_3$, $q_4$
  - $u$ (un-)
  - $e$ (adj-root)
  - $e$ (adj-root)
  - $e$ (adj-root)

$adj\text{-}root_1$: {happy, real, ...}
$adj\text{-}root_2$: {big, small, ...}

**FSA: Derivational Morphology**

- $q_0$, $q_1$, $q_2$, $q_3$, $q_4$, $q_5$
  - $verb$
  - $verb$
  - $verb$

**Agenda**

- HW1 – online tonight, due next Thursday
- Morphology
  - Words
  - Concatenative vs. non-concatenative
  - Inflectional vs. derivational morphology
  - Regular vs. irregular
  - Formal morphology
- Computational morphology
  - Finite-state methods: FSAs, FSTs