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-wir-a-nuhwe'-s
MsS-baby-∅-like-HAB
'He likes babies'

Alaskan Yupik (Woodbury 1987):

qayá:liyú:lú:ní
'he was excellent (-*yu*-) at making (-lí-) kayaks (qaya:-)'

Turkish (Hankamer, 1986):

cöplüklerimizdekileerdenmiydi
(garbage+AFF+PL+1P/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
  - Denominal adjectives
  - Denominal 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
  - hop+ing → hopping
- Affixes:
  - Prefixes: Anti**dis****establish**mentarianism
  - Suffixes: Antidis**establish**mentarianism
- Agglutinative languages (e.g., Turkish)
  - uygarlaştıramadıklarımızdanmışsınızcasına → uygar+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına
  - 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)
  - h**um**ingi (borrower)
- Circumfixes (e.g., German)
  - sagen (say)
  - ge**sag**t (said)
- Reduplication (e.g., Motu, spoken in Papua New Guinea)
  - mahuta (to sleep)
  - mahutamahuta (to sleep constantly)
  - mamahuta (to sleep, plural)

Infixation Examples

- (vs prefixation/suffixation of concatenative morphology)

Bontoc (Fromkin and Rodman 1983): Do normal affixation, ignoring a segment:

fikas	strong	fumikas	'be strong'
kilad	red	kumilad	'be red'
fusul	enemy	fumusul	'be an enemy'

Ulwa (CODIUL 1989) Parse out a prosodic unit–here a foot–and attach to it:

bilam	bilamki	'fish'/'my fish'
dii	diikimuih	'snake'/'my snake'
liima	liikima	'lemon'/'my lemon'
sikbilh	sikkibilh	'horsefly'/'my horsefly'

English: attach after a foot:

absolutely	abso-f******-lutely
Kalamazoo	Kalama-f******-zoo

Circumfixation Examples

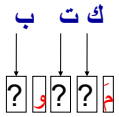
säuseln	**'rustle'**	**gesäuselt**	**'rustled'**
brüsten	**'brag'**	**gebrüstet**	**'bragged'**
täuschen	**'deceive'**	**getäuscht**	**'deceived'**

- 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

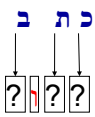
Arabic



مكتوب

maktuub
written

Hebrew



כתוב

ktuuv
written

More Templatic Morphology

Binyan	active	passive	template	gloss
I	katab	kutib	CVCVC	'write'
II	kattab	kuttib	CVCCVC	'cause to write'
III	kaatab	kuutib	CVVCVC	'correspond'
VI	takaatab	tukuutib	tVCVVCVC	'write to each other'
VII	nkaatab	nkuutib	nCVVCVC	'subscribe'
VIII	ktatab	ktutib	CtVCVC	'write'
X	staktab	stuktib	stVCCVC	'dictate'

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 English

Morphological Class	Regularly Inflected Verbs			
stem	walk	merge	try	map
-s form	walks	merges	tries	maps
-ing participle	walking	merging	trying	mapping
Past form or *-ed* participle	walked	merged	tried	mapped

Morphological Class	Irregularly Inflected Verbs		
stem	eat	catch	cut
-s form	eats	catches	cuts
-ing participle	eating	catching	cutting
preterite	ate	caught	cut
past participle	eaten	caught	cut

Verb Inflections in Spanish

	Present Indicative	Imperfect Indicative	Future	Preterite	Present Subjunctive	Conditional	Imperfect Subjunctive	Future Subjunctive
1SG	amo	amaba	amaré	amé	ame	amaría	amara	amare
2SG	amas	amabas	amarás	amaste	ames	amarías	amaras	amares
3SG	ama	amaba	amará	amó	ame	amaría	amara	amáreme
1PL	amamos	amábamos	amaremos	amamos	amemos	amaríamos	amáramos	amáremos
2PL	amáis	amabais	amaréis	amasteis	améis	amaríais	amarais	amareis
3PL	aman	amaban	amarán	amaron	amen	amarían	amaran	amaren

Derivational Morphology

- Stem + morpheme →
  - Word with different meaning or different part of speech
  - Exact meaning difficult to predict
- Nominalization in English:
  - -ation: computerization, characterization
  - -ee: appointee, advisee
  - -er: killer, helper
- Adjective formation in English:
  - -al: computational, derivational
  - -less: clueless, helpless
  - -able: teachable, computable

Examples of Derivational Morphology

- Agentive Nominals
  - adder, baker, catcher, dealer, eater, fighter, grinder, hater, ionizer, jumper, killer, lover, manager, namer, 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, pseudointellectial; semi-arid, semidivine, semiregular

Compound Morphology

- firefighter, football, firecracker, policeman, doghouse
- Lebensversicherungsgesellschaftsangestellter 'life insurance company employee'
- computer communications network performance analysis primer

Formal Morphology

- How is information formally encoded morphologically?
  - prefixation, suffixation
  - infixation
  - circumfixation
  - 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

WORD	**STEM (+FEATURES)***
cats	cat +N +PL
cat	cat +N +SG
cities	city +N +PL
geese	goose +N +PL
ducks	(duck +N +PL) or (duck +V +3SG)
merging	merge +V +PRES-PART
caught	(catch +V +PAST-PART) or (catch +V +PAST)

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 $V+s$
acclamation      acclamation $N$
acclamations     acclamation $N+s$
acclimate        acclimate   $V+0$
acclimated       acclimate   $V+ed$
acclimated       acclimate   $V+en$
acclimates       acclimate   $V+s$
acclimating      acclimate   $V+ing$
```


Rule-only: Porter Stemmer

- Cascading set of rules
  - ational → ate (e.g., relational → relate)
  - ing → ε (e.g., walking → walk)
  - sses → ss (e.g., grasses → grass)
  - …
- Examples
  - cities → citi
  - city→ citi
  - generalizations
    → generalization
    → generalize
    → general
    → gener

Porter Stemmer: What's the Problem?

- Errors…

Errors of Commission		Errors of Omission	
organization	organ	European	Europe
doing	doe	analysis	analyzes
numerical	numerous	noise	noisy
policy	police	sparse	sparsity

- Why is it still useful?

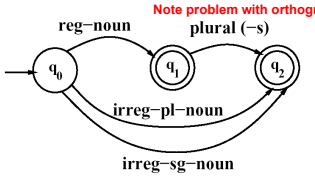
Lexicon + Rules

- 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

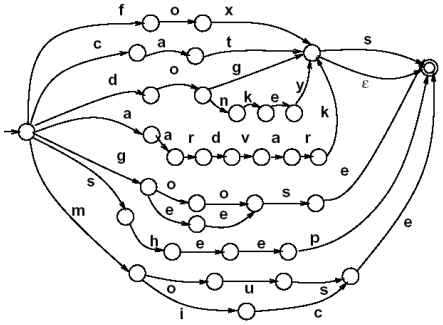
FSA: English Noun Morphology

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

Note problem with orthography!

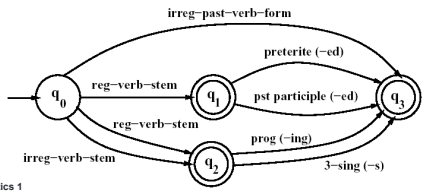


FSA: English Noun Morphology



FSA: English Verb Morphology

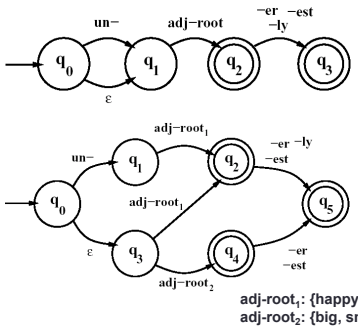
reg-verb-stem	irreg-verb-stem	irreg-past-verb	past	past-part	pres-part	3sg
walk fry talk impeach	cut speak spoken sing sang	caught ate eaten	-ed	-ed	-ing	-s



FSA: English Adjectival Morphology

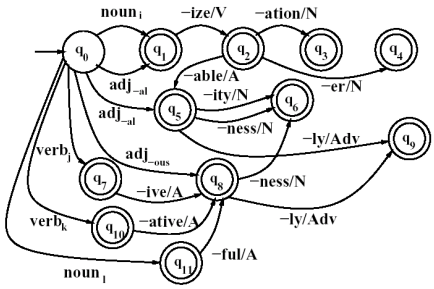
- Examples:
  - big, bigger, biggest
  - smaller, smaller, smallest
  - happy, happier, happiest, happily
  - unhappy, unhappier, unhappiest, unhappily
- Morphemes:
  - Roots: big, small, happy, etc.
  - Affixes: un-, -er, -est, -ly

FSA: English Adjectival Morphology



adj-root$_1$: {happy, real, …}
adj-root$_2$: {big, small, …}

FSA: Derivational Morphology



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