MORPHOLOGY AND THE FST

NATURAL LANGUAGE PROCESSING - CS 322.00

Blake Howald

September 18, 2017

AGENDA

- Questions, Logistics
- Morphology Generally
- Morphology of English
- Finite State Transducers (FST)

MORPHEMES AND WORD

- Phoneme – meaning-bearing unit at the phonetic level: ***f****ox* - ***b****ox*
- Morpheme – meaning-bearing unit at the multi-phoneme/ "word" level: *box* – *box* + ***es*** (one vs. two morphemes)

  Box = **stem** : *es* = **affix**

  Affixation – Prefix, Suffix, Infixes and Circumfixes
- Linguistic evidence for "words":
  - Positional Mobility: *I love* ***coffee****.* | ***Coffee****, I love it.* | *****Cof,*** *I love it* ***fee****.*
  - Internal Stability: *Coffee* is always 'coffee', never ***Fcoeef***

MORPHOLOGICAL TYPOLOGY

- Languages pattern in terms of (1) the degree of *affixation* permitted in the language and (2) the word to morpheme ratio
- Despite clear examples of each type, a given language tends to be mixed (show characteristics of several types)
  - Analytic/Isolating Languages
    - Analytic Languages characterized by a one to one, word to morpheme correspondence, with some affixation (and compounding) permitted.
    - Isolating languages are the extreme cases – purely analytic, no affixation. In Mandarin, for example:

      我們彈鋼琴

      [wɔ mən tan tɕin lə]

      1P +Pl play piano +Past

      '*we played the piano*'

MORPHOLOGICAL TYPOLOGY

- Synthetic Languages are characterized by a one to two (or more) word to morpheme correspondence:
  - ***Agglutinative*** – clear delineation of morphemes (even if one word to many) and morphemes have a one to one correspondence of meaning – e.g., in Turkish

    el-ler-imiz-in

    'hand' – +Pl – 1stPerson+Pl – genitive case

    'of our hands'
  - ***Fusional*** – morphemes may have a one to many correspondence of meaning – e.g, in Spanish

    abl-o

    'speak' – 1stPerson+Present Tense

    'I am speaking'
  - ***Polysynthetic*** – high degree of affixation – West Greenlandic

    tusaa-nngit-su-usaar-tuaannar-sinnaa-nngi-vip-putit

    'hear'-neg.-intrans.participle-'pretend'-'all the time'-'can'-neg.-''really'-2nd.sng.indicative

    '*You simply cannot pretend not to be hearing all the time'*

MORPHOLOGICAL CONCEPTS

- Combining morphemes:
  - Inflection – Stem + Grammatical Morpheme ***with no change*** in the class of the Stem: plurals, possessives, conjugations, etc.
  - Derivation – Stem + Grammatical Morpheme ***with a change*** in the class of the Stem.
  - Compounding – Stem + Stem(s) for larger words: *doghouse, carpark*
  - Cliticization – a word that has been phonologically (and orthographically) altered to be combined with another word: *I've = I have.*

IN ENGLISH

- Eight inflectional affixes of English (not irregulars):
  - Nouns
    - Plural *–s*
    - Possessive *–s*
  - Adjectives
    - Comparative *–er*
    - Superlative *–est*
  - Verbs
    - Past Tense *–ed*
    - Past Participle *–en*
    - 3rd Person Present Tense *–s*
    - Present Participle *–ing*

IN ENGLISH

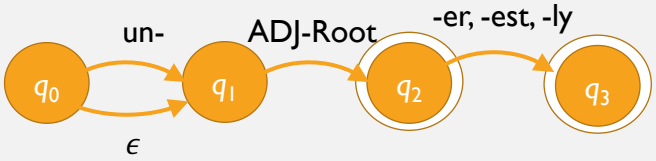
- Examples of derivational affixes:
  - Verb/Adjective => Noun: *-er*, *-ness*, *-ation*
  - Noun/Verb => Adjective: *-less*, *-able*
- Examples of cliticization:
  - Had, Would – *'d*
  - Has, Is –*'s*
  - Am – *'m*
  - Like affixes, clitics can occur before and after a word (proclitic and enclitic, respectively)

ADDITIONAL CONSIDERATIONS

- Agreement
  - Gender, Number, Case, Tense, etc.
- Templative Morphology
  - Common in Semitic languages where there is some root – CCC which is varied to create meaning
- Parsing Morphology means to break a candidate string into morphologically salient components:
  - 'goose' – goose + N + Sg
  - 'geese' – goose + N + Pl
  - 'goose' – goose +V
  - 'gooses' – goose +V +3P +Sg

MORPHOLOGY WITH FSAS

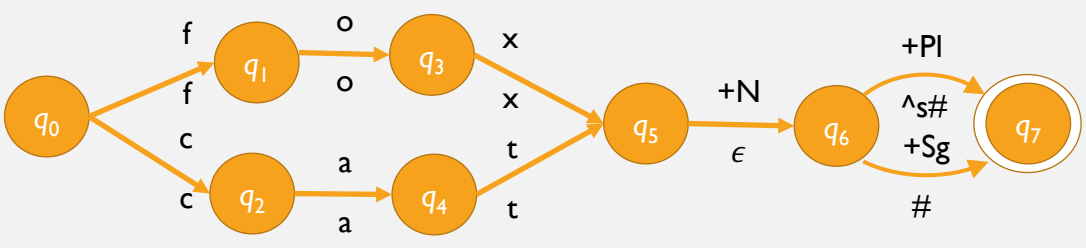
- You need some things:
  - ***Lexicon*** – all of the stems, affixes and corresponding grammatical information associated with input
  - ***Morphotactics*** – modeling of the interactions between the stems and affixes associated with input -e.g. superlative –est can only go on the ends of adjectives
  - ***Orthographic Rules*** – accounting for any surface variations and changes in the input that may obscure mappings
- FSAs can be altered into a Finite State Transducer (FST) to model the Morphotactics (and Orthographic Rules).



FST

- A Two-Tape (Lexical and Surface) FSA with some additions and modifications to the definition of a traditional FSA:
  - Output Alphabet (Δ)
  - Output Function: $\sigma(q,w)$
    - Given $q \in Q$ and $w \in \Sigma$, $\delta(q,w)$ returns $o \in \Delta$
  - w is a string of input symbols from the input alphabet ($i \in \Sigma$) or output alphabet ($o \in \Delta$)
  - Symbols in alphabet are paired:
    - Σ {a:a, r:r, g:g, h:h, !:!, ε:r, +N: ε, +Pl:^s#}
      - Default pairs are symbols that map to themselves
      - Include grammatical morphemes (+N, +Pl), morpheme boundaries (^) and word boundaries (#)

FST



FST – ORTHOGRAPHIC RULES

- Additional Intermediate Tape between Lexical and Surface Tapes
  - FST for Lexical to Intermediate, FST for Intermediate to Surface
- FST from Intermediate to Surface takes the form or rewrite rules:

cat + N:ϵ + Pl:^s#	fox + N:ϵ + Pl:^s#
cat + ϵ:ϵ + ^s#:s	fox + ϵ:e + ^s#:s
cats	foxes

FST ADDITIONAL CONSIDERATIONS

- FSTs can be inverted to run backwards (swapping input and output labels)
- FSTs can be concatenated – run parse and orthography in one (albeit complex) FST
- Ambiguities have a bigger impact on parsing than generation.
- Stemming vs. Lemmatizing
  - Stemming is the stripping off of endings to get roots/stems (singer, sings, singing: stem = sing
    - The Porter Stemmer is worth looking at (https://tartarus.org/martin/PorterStemmer/def.txt)
  - Lemmatizing is finding the common root between multiple words (sing, sang, sung: lemma = sing)