CSCI 5832
Natural Language Processing

Lecture 4

Jim Martin

Today 1/25

- **More English Morphology**
- **FSAs and Morphology**
- **Break**
- **FSTs**

English Morphology

- **Morphology is the study of the ways that words are built up from smaller meaningful units called morphemes**
- **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**

Inflectional Morphology

- **Inflectional morphology concerns the combination of stems and affixes where the resulting word**
  - **Has the same word class as the original**
  - **Serves a grammatical/semantic purpose different from the original**

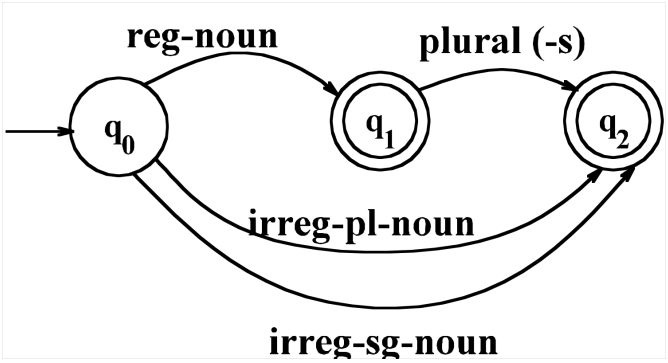
Nouns and Verbs (English)

- **Nouns are simple (not really)**
  - **Markers for plural and possessive**
- **Verbs are only slightly more complex**
  - **Markers appropriate to the tense of the verb**

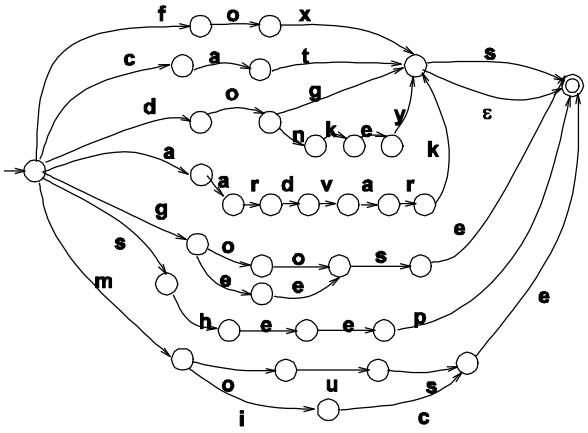
FSAs and the Lexicon

- **First we'll capture the morphotactics**
  - **The rules governing the ordering of affixes in a language.**
- **Then we'll add in the actual words**

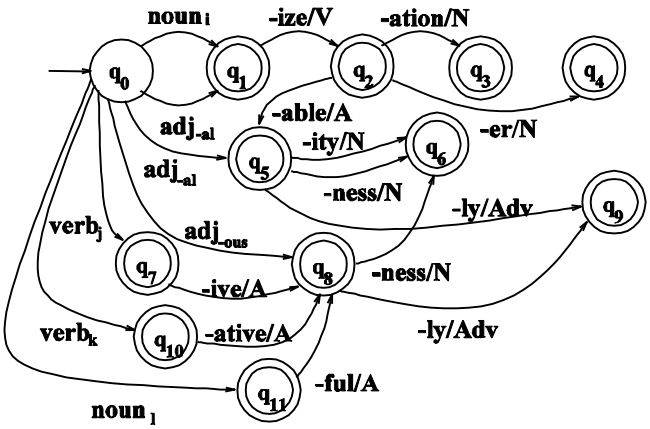
Simple Rules



Adding the Words



Derivational Rules



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**

Homework

- **How big is your vocabulary?**

Projects

- **2 styles of projects**
  - **Something no one has done…**
    - **You might ask yourself why no one has done it.**
  - **Tasks that have benchmarks and current best results from bakeoffs**
- **To get ideas about the latter go to acl.ldc.upenn.edu and poke around.**

Projects

- **Other ideas…**
  - **Anything to do with blogs**
  - **Machine learning applied to X**
    - **Clustering (unsupervised)**
    - **Classification (supervised)**
  - **Bioinformatic language sources**
  - **Search engines (getting old)**
  - **Semantic tagging (getting hot)**

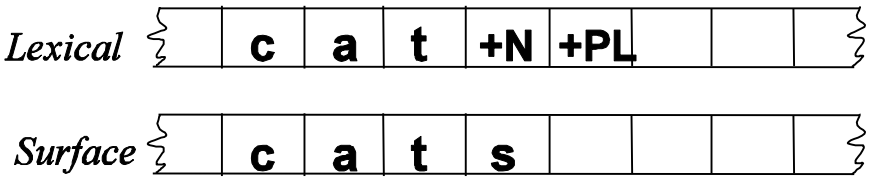
Applications

- **The kind of parsing we're talking about is normally called morphological analysis**
- **It can either be**
  - **An important stand-alone component of an application (spelling correction, information retrieval)**
  - **Or simply a link in a chain of processing**

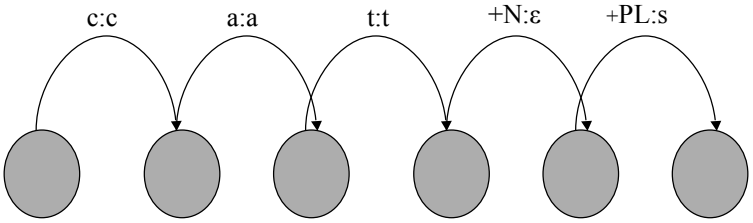
Finite State Transducers

- **The simple story**
  - **Add another tape**
  - **Add extra symbols to the transitions**
  - **On one tape we read "`cats`", on the other we write "`cat +N +PL`", or the other way around.**

FSTs



Transitions



- **c:c means read a c on one tape and write a c on the other**
- **+N:ε means read a +N symbol on one tape and write nothing on the other**
- **+PL:s means read +PL and write an s**

Typical Uses

- **Typically, we'll read from one tape using the first symbol on the machine transitions (just as in a simple FSA).**
- **And we'll write to the second tape using the other symbols on the transitions.**

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.

Ambiguity

- 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, its 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

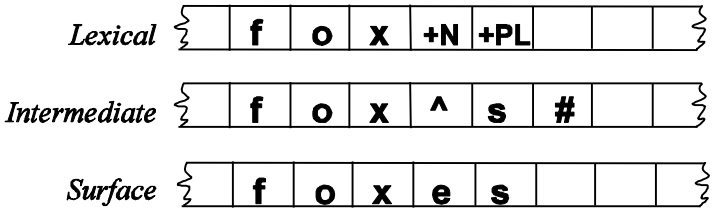
Multi-Tape Machines

- To deal with this we can simply add more tapes and use the output of one tape machine as the input to the next
- So to handle irregular spelling changes we'll add intermediate tapes with intermediate symbols

Generativity

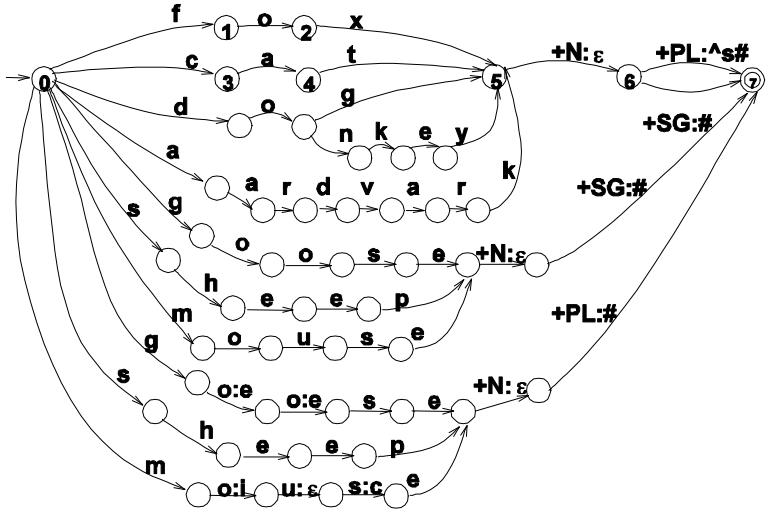
- Nothing really privileged about the directions.
- We can write from one and read from the other or vice-versa.
- One way is generation, the other way is analysis

Multi-Level Tape Machines



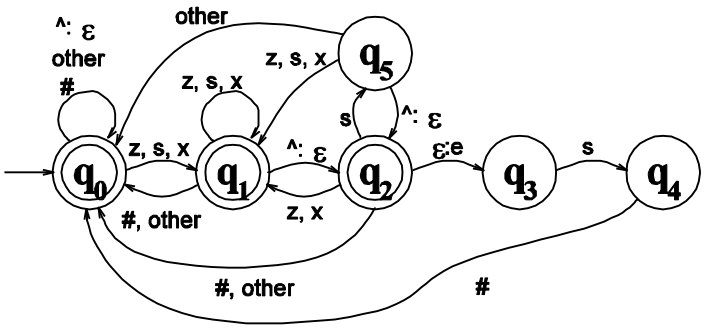
- **We use one machine to transduce between the lexical and the intermediate level, and another to handle the spelling changes to the surface tape**

Lexical to Intermediate Level

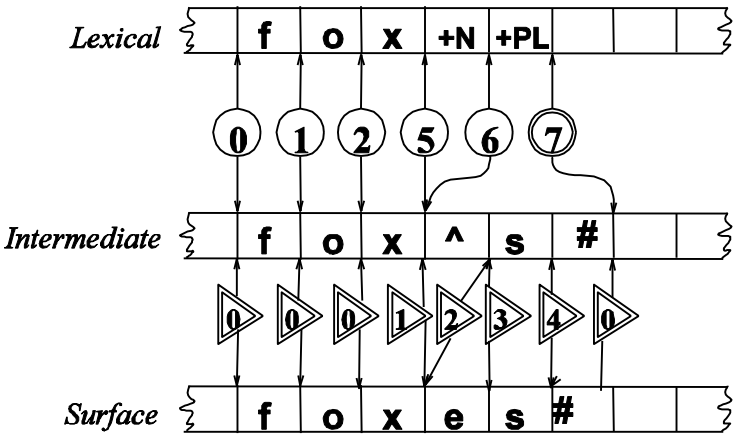


Intermediate to Surface

- **The add an "e" rule as in `fox^s# <-> foxes#`**



Foxes



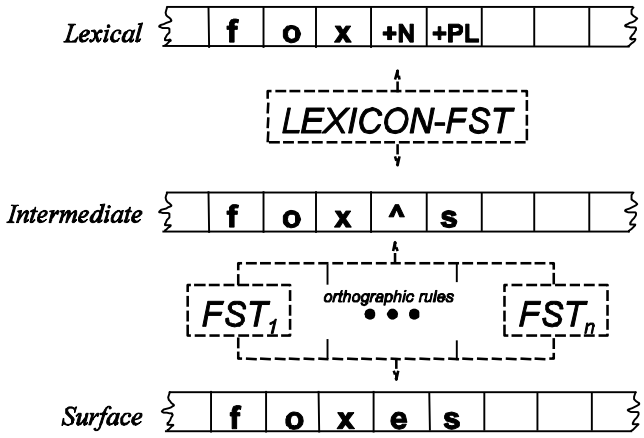
Note

- 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



Next Time

- **Finish Chapter 3**