CSCI 5832
Natural Language Processing

Lecture 5
Jim Martin

Today 1/31

- **Review**
- **FSTs/Composing cascades**
- **Administration**
- **FST Examples**
  - **Porter Stemmer**
  - **Soundex**

FST Review

- FSTs allow us to take an input and deliver a structure based on it
- Or… take a structure and create a surface form
- Or take a structure and create another structure

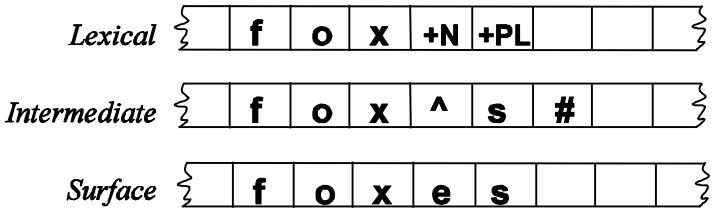
FST Review

- What does the transition table for an FSA consist of?
- How does that change with an FST?

Review

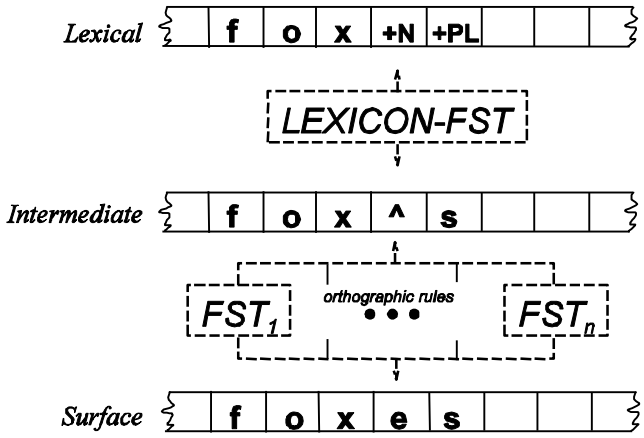
- **In many applications its convenient to decompose the problem into a set of cascaded transducers where**
  - **The output of one feeds into the input of the next.**
  - **We'll see this scheme again for deeper semantic processing.**

English Spelling Changes

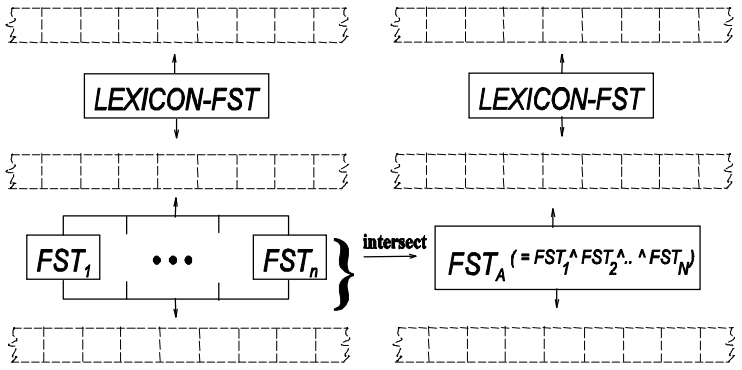


- **We use one machine to transduce between the lexical and the intermediate level, and another to handle the spelling changes to the surface tape**

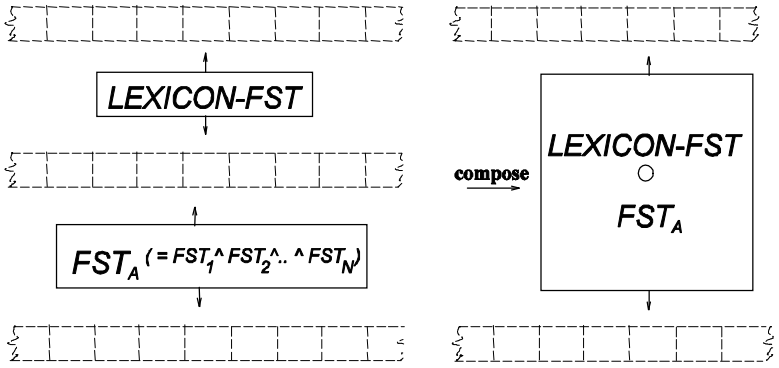
Overall Plan



Final Scheme: Part 1



Final Scheme: Part 2



Composition

1. **Create a set of new states that correspond to each pair of states from the original machines (New states are called (x,y), where x is a state from M1, and y is a state from M2)**
2. **Create a new FST transition table for the new machine according to the following intuition…**

Composition

- There should be a transition between two related states in the new machine if it's the case that the output for a transition from a state from M1, is the same as the input to a transition from M2 or…

Composition

- $\delta_3((x_a,y_a), i:o) = (x_b,y_b)$ iff
  - There exists c such that
  - $\delta_1(x_a, i:c) = x_b$ AND
  - $\delta_2(y_a, c:o) = y_b$

Extra HW Post-Mortem

- **What does the NY Times think a word is?**

Sentence Segmentation

- **Titles**
  - **Mr. Mrs. Ms.**
  - **Just bigger lists? Or something better?**
- **Initials in names**
  - **H. W.**
  - **Regular expressions?**
- **Numbers**
  - **$2.8 million**
- **Acronyms**
  - **P.O.W., I.R.A., …**
  - **Lists? Regexs?**

Sentence Segmentation

- **This could get tedious…**
  - **Lots of rules, lots of exceptions, brittle**
- **Another approach is to use machine learning… Take a segmented corpus of text and learn to segment it.**
- **Not clear if it is better in this case, but it is the dominant approach these days.**

Next HW

- **Part 1: Alter your code to take newswire text and segment it into paragraphs, sentences and words. As in…**

```
<p>
<s> The quick brown fox. </s>
<s> This is sentence two. </s>
</p>
…
```

Projects

- Read ahead in the book to get a feel for various areas of NLP
- The goal is to produce a paper that could be sent to a conference on NLP
- To get a feel for what that means you need to get familiar with such papers
  - acl.ldc.upenn.edu

ACL Repository

- Lots of stuff: all relevant, not all terribly readable
- But focus on recent
  - ACL conference proceedings
  - NAACL proceedings
  - HLT proceedings
- Just start browsing titles that look interesting.

Readings/Quiz

- First quiz is 2/8 (a week from Thursday).
- It will cover Chapters 2,3,4 and part of 5.
- Lectures are based on the assumption that you've read the text before class.
- Quizzes are based on the contents of the lectures and the chapters.

Light Weight Morphology

- Sometimes you just need to know the stem of a word and you don't care about the structure.
- In fact you may not even care if you get the right stem, as long as you get a consistent string.
- This is stemming… it most often shows up in IR applications

Stemming for Information Retrieval

- Run a stemmer on the documents to be indexed
- Run a stemmer on users' queries
- Match
  - This is basically a form of hashing, where you want collisions.

Porter

- No lexicon needed
- Basically a set of staged sets of rewrite rules that strip suffixes
- Handles both inflectional and derivational suffixes
- Doesn't guarantee that the resulting stem is really a stem (see first bullet)
- Lack of guarantee doesn't matter for IR

Porter Example

- **Computerization**
  - **ization -> -ize computerize**
  - **ize -> ϵ computer**

Porter

- **The original exposition of the Porter stemmer did not describe it as a transducer but…**
  - **Each stage is separate transducer**
  - **The stages can be composed to get one big transducer**

Soundex

- **You work as a telephone information operator for CU. Someone calls looking for our senior theory professor…**
  - **Ehrenfeucht**
  - **What do you type as your query string?**

Soundex

1. **Keep the first letter**
2. **Drop non-initial occurrences of vowels, h, w and y**
3. **Replace the remaining letters with numbers according to group (e.g.. b, f, p, and v -> 1**
4. **Replace strings of identical numbers with a single number (333 -> 3)**
5. **Drop any numbers beyond a third one**

Soundex

- **Effect is to map (hash) all similar sounding transcriptions to the same code.**
- **Structure your directory so that it can be accessed by code as well as by correct spelling**
- **Used for census records, phone directories, author searches in libraries etc.**

Next Time

On to Chapter 4 for Thursday.