

Words: **Tokenization** **Spelling**

Slides by James Martin, adapted by Diana Inkpen
for CSI 5386 @ uOttawa

Segmentation

- ◆ Lightweight morphology (stemming)
- ◆ Tokenization
- ◆ Sentence segmentation
- ◆ Spell checking/correction
- ◆ Edit distance

Tokenizing

- Identifying the tokens (words) in a text that we may want to deal with
- Pretty much a prerequisite to doing anything interesting
- Difficulty varies by genre, task and language

Tokenizing

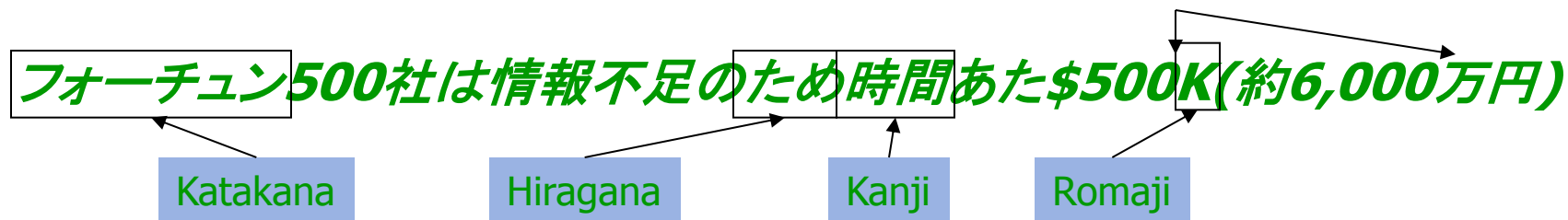
- For English, why not just use white-space?
 - ♦ Mr. Sherwood said reaction to Sea Containers' proposal has been "very positive." In New York Stock Exchange composite trading yesterday, Sea Containers closed at \$62.625, up 62.5 cents.
 - ♦ "I said, 'what're you? Crazy?' " said Sadowsky. "I can't afford to do that.'"
- Using white-space gives you words like:
 - ♦ cents.
 - ♦ said,
 - ♦ positive."
 - ♦ Crazy?

Punctuation Issues

- Word-internal punctuation
 - ♦ M.P.H.
 - ♦ Ph.D.
 - ♦ AT&T
 - ♦ 01/02/06
 - ♦ Google.com
 - ♦ Yahoo!
 - ♦ 555,500.50
- Clitics
 - ♦ What're
 - ♦ I'm
- Multi-token words
 - ♦ New York
 - ♦ Rock 'n' roll
 - ♦ Ice Capades

Language Issues

- Chinese has no spaces between words
 - ◆ 莎拉波娃现在居住在美国东南部的佛罗里达。
 - ◆ 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
 - ◆ Sharapova now lives in US southeastern Florida
- Further complicated in languages that allow multiple alphabets to be intermingled
 - ◆ Dates/amounts in multiple formats



Slide from Chris Manning

Segmentation in Chinese

- Words composed of characters
- Characters are generally 1 syllable and 1 morpheme.
- Average word is 2.4 characters long.
- Standard segmentation algorithm:
 - ◆ Maximum Matching or Maxmatch

Maximum Matching Word Segmentation

Given a lexicon of Chinese, and a string

- 1) Start a pointer at the beginning of the string
- 2) Find the longest word in dictionary that matches the string starting at pointer
 - 1) If there are no matches, emit a character and advance the pointer 1 character
- 3) Move the pointer over the word in string
- 4) Go to 2

English Example

thetabledownthere

theta bled own there the table down there

- But works pretty well in Chinese
 - ◆ 莎拉波娃现在居住在美国东南部的佛罗里达。
 - ◆ 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- We'll see better ones soon based on probabilities

Practical English Segmentation Examples

- URL segmentation
 - ◆ www.dietsthatwork.com
 - ◆ www.choosespain.com
- Twitter hashtag segmentation
 - ◆ #unitedbrokemyguitar
 - ◆ #manchesterunited

Spelling Correction

- We can **detect** spelling errors (spell check) by building an FST-based lexicon and noting any strings that are rejected.
- But how do I fix “**graffe**”? That is, how do I come up with suggested corrections?
 - ◆ Search through all words in my lexicon
 - Graft, craft, grail, giraffe, crafted, etc.
 - ◆ Pick the one that’s closest to **graffe**
 - ◆ But what does “closest” mean?
 - We need a **distance metric**.
 - The simplest one: **minimum edit distance**
 - Ala Unix diff

Edit Distance

- The minimum edit distance between two strings is the minimum number of editing operations
 - ◆ Insertion
 - ◆ Deletion
 - ◆ Substitution

that one would need to transform one string into the other

Note

- The following discussion has 2 goals
 1. Introduce the minimum edit distance computation and algorithm
 2. Introduce dynamic programming

Why “Dynamic Programming”

“I spent the Fall quarter (of 1950) at RAND. My first task was to find a name for multistage decision processes. An interesting question is, Where did the name, dynamic programming, come from? **The 1950s were not good years for mathematical research. We had a very interesting gentleman in Washington named Wilson. He was Secretary of Defense, and he actually had a pathological fear and hatred of the word, research.** I’m not using the term lightly; I’m using it precisely. His face would suffuse, he would turn red, and **he would get violent if people used the term, research, in his presence.** You can imagine how he felt, then, about the term, mathematical. The RAND Corporation was employed by the Air Force, and the Air Force had Wilson as its boss, essentially. Hence, I felt I had to do something to shield Wilson and the Air Force from the fact that I was really doing mathematics inside the RAND Corporation. What title, what name, could I choose? In the first place I was interested in planning, in decision making, in thinking. But planning, is not a good word for various reasons. **I decided therefore to use the word, “programming” I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying I thought, lets kill two birds with one stone. Lets take a word that has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that is its impossible to use the word, dynamic, in a pejorative sense.** Try thinking of some combination that will possibly give it a pejorative meaning. Its impossible. Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities.”

Richard Bellman, “Eye of the Hurricane: an autobiography” 1984.



Min Edit Example

	i	n	t	e	n	t	i	o	n	
delete i →		n	t	e	n	t	i	o	n	
substitute n by e →		e	t	e	n	t	i	o	n	
substitute t by x →		e	x	e	n	t	i	o	n	
insert u →		e	x	e	n	u	t	i	o	n
substitute n by c →		e	x	e	c	u	t	i	o	n

Minimum Edit Distance

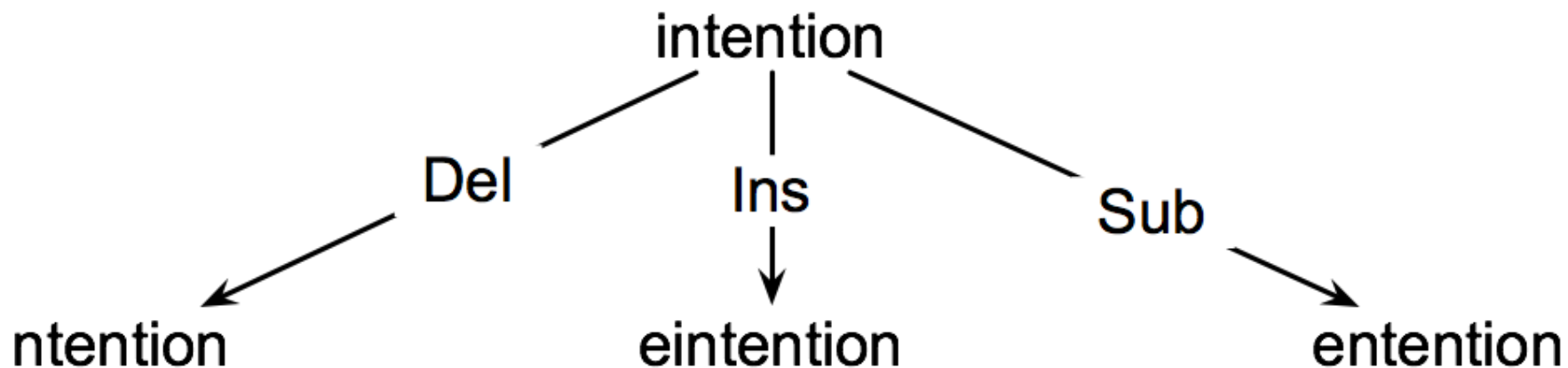
I N T E * N T I O N
| | | | | | | | |
* E X E C U T I O N
d s s i s

- If each operation has cost of 1
- Distance between these is 5
- If substitutions cost 2 (Levenshtein)
- Distance between these is 8

Min Edit As Search

- That's all well and good but how did we find that particular (minimum) set of operations for those two strings?
 - ◆ As opposed to say, deleting all the characters in one and then inserting the characters of the other (18)
- We can view edit distance as a search for a path (a sequence of edits) that gets us from the start string to the final string
 - ◆ Initial state is the word we're transforming
 - ◆ Operators are insert, delete, substitute
 - ◆ Goal state is the word we're trying to get to
 - ◆ Path cost is what we're trying to minimize: the number of edits

Min Edit as Search



Min Edit As Search

- But that generates a huge search space
- Navigating that space in a naïve backtracking fashion would be incredibly wasteful
- Why?

Many distinct paths (sequence of edits) wind up at the same intermediate states. But there is no need to keep track of the them all. We only care about the shortest path to each of those revisited states.

Defining Min Edit Distance

- For two strings S_1 of len n , S_2 of len m
 - ◆ $\text{distance}(i,j)$ or $D(i,j)$
 - means the edit distance of $S_1[1..i]$ and $S_2[1..j]$
 - i.e., the minimum number of edit operations need to transform the first i characters of S_1 into the first j characters of S_2
 - The edit distance of S_1, S_2 is $D(n,m)$
- We compute $D(n,m)$ by computing $D(i,j)$ for all i ($0 < i < n$) and j ($0 < j < m$)

Defining Min Edit Distance

- Base conditions:

- ◆ $D(i,0) = i$

- ◆ $D(0,j) = j$

- ◆ Recurrence Relation:

- ◆ $D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$

Dynamic Programming

- A tabular computation of $D(n,m)$
- Bottom-up
 - ◆ We compute $D(i,j)$ for small i,j
 - ◆ And compute larger $D(i,j)$ based on previously computed smaller values

The Edit Distance Table

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Min Edit Distance

- Note that the result isn't all that informative
 - ◆ For a pair of strings we get back a single number
 - The min number of edits to get from here to there
- That's sort of like a map routing program that tells you the distance from here to Crested Butte but doesn't tell you how to get there.

Alignment

- An alignment is a 1 to 1 pairing of each element in a sequence with a corresponding element in the other sequence or with a gap...

```

I N T E * N T I O N
| | | | | | | | |
* E X E C U T I O N
d s s   i s

```

```

-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTGCCCCGAC

```

Paths / Alignments

- Keep a back pointer
 - ◆ Every time we fill a cell add a pointer back to the cell that was used to create it (the min cell that lead to it)
 - ◆ To get the sequence of operations follow the backpointer from the final cell
 - ◆ That's the same as the alignment.

Backtrace

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Adding Backtrace to MinEdit

- Base conditions:

- ♦ $D(i,0) = i$
- ♦ $D(0,j) = j$

- Recurrence Relation:

$$\diamond D(i,j) = \min \left\{ \begin{array}{l} D(i-1,j) + 1 \quad \text{Case 1} \\ D(i,j-1) + 1 \quad \text{Case 2} \\ D(i-1,j-1) + \begin{cases} 1; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \quad \text{Case 3} \end{array} \right.$$

$$\text{ptr}(i,j) \left\{ \begin{array}{l} \text{LEFT} \quad \text{Case 1} \\ \text{DOWN} \quad \text{Case 2} \\ \text{DIAG} \quad \text{Case 3} \end{array} \right.$$

Complexity

- Time:

$$O(nm)$$

- Space:

$$O(nm)$$

- Backtrace

$$O(n+m)$$

DP Search

- In the context of language processing (and signal processing) this kind of algorithm is often referred to as a DP search
 - ◆ Min edit distance
 - ◆ Viterbi and Forward algorithms
 - ◆ CKY and Earley
 - ◆ MT decoding