LECTURE 43: EVALUATION METRICS

- Objectives:
  - Human Performance
  - Machine Performance
  - Automated Scoring: String Alignment
  - Precision and Recall

This lecture uses material from the instructor's notes. Most NLP books contain information about scoring. A good resource is:

D. Jurafsky and J.H. Martin,

SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall, ISBN: 0-13-095069-6, 2000.

LECTURE 43: EVALUATION METRICS

- Objectives:
  - Human Performance
  - Machine Performance
  - Automated Scoring: String Alignment
  - Precision and Recall

This lecture uses material from the instructor's notes. Most NLP books contain information about scoring. A good resource is:

D. Jurafsky and J.H. Martin, *SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing*,

Computational Linguistics, and Speech Recognition, Prentice-Hall, ISBN: 0-13-095069-6, 2000.

AUTOMATED SCORING USING STRING EDITS AND DYNAMIC PROGRAMMING

To automatically score a hypothesis, we must first align it with the reference text, and then count word errors (substitutions, deletions, and insertions).

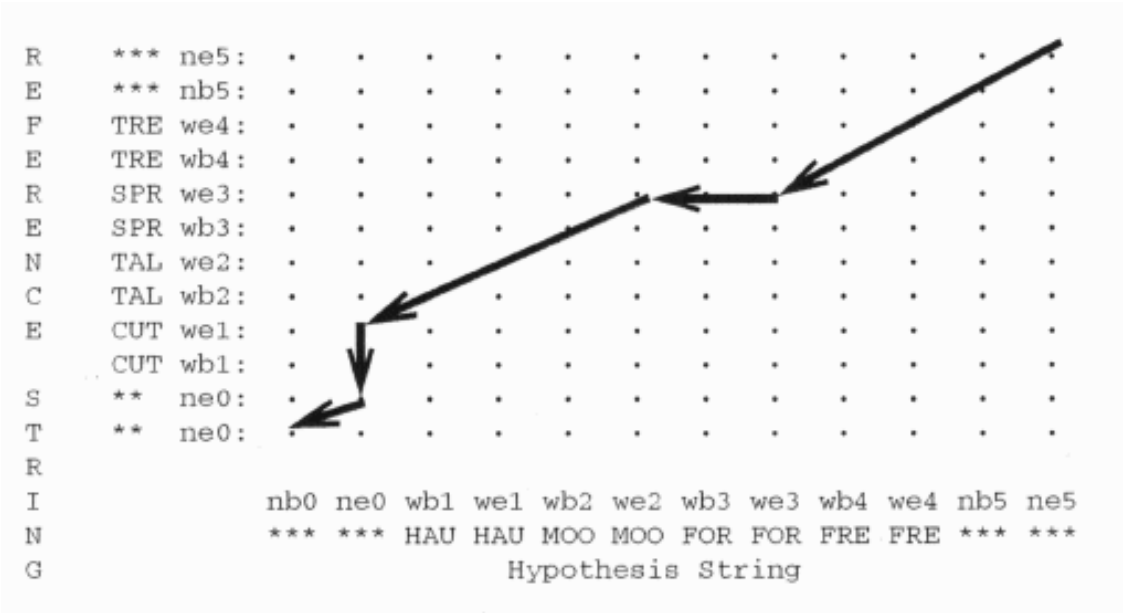
The desired output is shown below:

```
Input REF: CUT TALL SPRUCE TREES
Input HYP: HAUL MOOSE FOR FREE

Align REF: CUT TALL SPRUCE *** TREES
Align HYP: *** HAUL MOOSE  FOR FREE

< 3 Sub | 1 Ins | 1 Del | 0 Cor | 4 Ref Words >
```

The solution to this problem can be achieved using dynamic programming with a Levenstein distance metric (each non-matching pair adds one to the accumulated distance). We can demonstrate this using a DP grid:


REFERENCE STRING
*** ne5:
*** nb5:
TRE we4:
TRE wb4:
SPR we3:
SPR wb3:
TAL we2:
TAL wb2:
CUT we1:
CUT wb1:
** ne0:
** ne0:
nb0 ne0 wb1 we1 wb2 we2 wb3 we3 wb4 we4 nb5 ne5
*** *** HAU HAU MOO MOO FOR FOR FRE FRE *** ***
Hypothesis String

THE NIST SCORING REPORT

A typical scoring report from the NIST standard scoring software is shown below:

```
DETAILED OVERALL REPORT FOR THE SYSTEM:
hypotheses_808080_total.out

SENTENCE RECOGNITION PERFORMANCE

 sentences
12547
 with errors                                1.9%   (
241)

   with substitions                         1.1%   (
134)
   with deletions                           0.2%   (
20)
   with insertions                          0.8%   (
102)


WORD RECOGNITION PERFORMANCE

Percent Total Error       =    0.6%   ( 263)
```

```
Percent Correct           =   99.6%   (41061)

Percent Substitution      =    0.3%   ( 138)
Percent Deletions         =    0.1%   (  21)
Percent Insertions        =    0.3%   ( 104)
Percent Word Accuracy     =   99.4%


Ref. words                =           (41220)
Hyp. words                =           (41303)
Aligned words             =           (41324)

CONFUSION PAIRS                       Total
(38)
                                      With >=  1
occurances (38)

   1:   13  ->  five ==> oh
   2:   12  ->  oh ==> nine
   3:    9  ->  nine ==> oh
   4:    8  ->  two ==> three
   5:    7  ->  oh ==> eight
   6:    7  ->  oh ==> four
   7:    6  ->  four ==> five
   8:    5  ->  eight ==> three
```

```
 9:    5  ->  five ==> nine
10:    5  ->  four ==> oh
11:    5  ->  three ==> eight
12:    5  ->  zero ==> oh
13:    4  ->  oh ==> seven
14:    4  ->  seven ==> oh
15:    4  ->  three ==> two
16:    3  ->  eight ==> six
17:    3  ->  eight ==> two
18:    3  ->  nine ==> one
19:    3  ->  oh ==> two
20:    3  ->  two ==> oh
21:    2  ->  eight ==> one
22:    2  ->  five ==> eight
23:    2  ->  nine ==> five
24:    2  ->  oh ==> zero
25:    2  ->  seven ==> one
26:    2  ->  six ==> eight
27:    1  ->  eight ==> five
28:    1  ->  eight ==> nine
29:    1  ->  eight ==> seven
30:    1  ->  four ==> one
31:    1  ->  one ==> five
32:    1  ->  one ==> four
33:    1  ->  seven ==> nine
34:    1  ->  seven ==> six
```

```
  35:    1  ->  seven ==> zero
  36:    1  ->  six ==> three
  37:    1  ->  three ==> one
  38:    1  ->  zero ==> two
     -------
       138



INSERTIONS                          Total
(11)
                                    With >=  1
occurances (11)

   1:   43  ->  oh
   2:   17  ->  eight
   3:   13  ->  six
   4:    9  ->  one
   5:    8  ->  nine
   6:    6  ->  two
   7:    3  ->  three
   8:    2  ->  four
   9:    1  ->  five
  10:    1  ->  seven
  11:    1  ->  zero
     -------
```

104

DELETIONS (3)

Total With >= 1

occurances (3)

1:	11	->	oh
2:	6	->	eight
3:	4	->	two
	21		

SUBSTITUTIONS (11)

Total With >= 1

occurances (11)

1:	35	->	oh
2:	20	->	five
3:	16	->	eight
4:	14	->	nine

```
   5:   12  ->  four
   6:   11  ->  two
   7:   10  ->  three
   8:    9  ->  seven
   9:    6  ->  zero
  10:    3  ->  six
  11:    2  ->  one
     -------
       138


* NOTE: The 'Substitution' words are those
reference words
        for which the recognizer supplied an
incorrect word.


FALSELY RECOGNIZED                    Total
(11)
                                      With >=  1
occurances (11)

   1:   39  ->  oh
   2:   19  ->  nine
   3:   16  ->  eight
   4:   14  ->  three
```

```
    5:   11  ->  two
    6:   10  ->  five
    7:    9  ->  one
    8:    8  ->  four
    9:    5  ->  seven
   10:    4  ->  six
   11:    3  ->  zero
      -------
        138


* NOTE: The 'Falsely Recognized' words are those
hypothesis words
         which the recognizer incorrectly
substituted for a reference word.



                 DUMP OF SYSTEM ALIGNMENT STRUCTURE

System name:   hypotheses_808080_total.out

Speakers:
     0:  bg
     1:  bk
     ...
```

```
   161:  sn
   162:  tb

Speaker sentences   0:  bg   #utts: 77
id: (bg_119oo39a)
Scores: (#C #S #D #I) 7 0 0 0
REF:  one one nine oh oh three nine
HYP:  one one nine oh oh three nine
Eval:

id: (bt_41722a)
Scores: (#C #S #D #I) 5 0 0 1
REF:  four ** one seven two two
HYP:  four OH one seven two two
Eval:      I

id: (gf_886374oa)
Scores: (#C #S #D #I) 6 1 0 0
REF:  eight eight six three seven FOUR oh
HYP:  eight eight six three seven OH   oh
Eval:                             S

...
id: (gf_886a)
Scores: (#C #S #D #I) 3 0 0 0
REF:  eight eight six
```

```
HYP:  eight eight six
Eval:

id: (gf_892a)
Scores: (#C #S #D #I) 3 0 0 0
REF:  eight nine two
HYP:  eight nine two
Eval:

id: (gf_8a)
Scores: (#C #S #D #I) 1 0 0 0
REF:  eight
HYP:  eight
Eval:

id: (gf_8b)
Scores: (#C #S #D #I) 1 0 0 0
REF:  eight
HYP:  eight
Eval:

id: (gf_8o156a)
Scores: (#C #S #D #I) 5 0 0 0
REF:  eight oh one five six
HYP:  eight oh one five six
Eval:
```

```
id: (gf_914a)
Scores: (#C #S #D #I) 3 0 0 0


                    SYSTEM SUMMARY PERCENTAGES by
SPEAKER


,-----------------------------------------------------------.
       |
hypotheses_808080_total.out                                 |

|-----------------------------------------------------------|
       | SPKR   | # Snt # Wrd | Corr    Sub    Del
Ins    Err  S.Err |

|--------+--------------+-----------------------------------|
       | bg     |   77    253 |100.0    0.0    0.0
0.0    0.0    0.0 |

|--------+--------------+-----------------------------------|
       | bk     |   77    253 | 99.6    0.4    0.0
0.0    0.4    1.3 |

|--------+--------------+-----------------------------------|
```

```
        ...

|=============================================================|
        | Sum/Avg|12547  41220 | 99.6    0.3    0.1
0.3    0.6    1.9 |

|=============================================================|
        |  Mean  | 77.0  252.9 | 99.6    0.3    0.1
0.3    0.6    1.9 |
        |  S.D.  |  0.2    1.1 |  1.1    1.0    0.2
0.7    1.4    3.9 |
        | Median | 77.0  253.0 |100.0    0.0    0.0
0.0    0.4    1.3 |

`-------------------------------------------------------------'

                     SYSTEM SUMMARY PERCENTAGES by
SPEAKER

,-------------------------------------------------------------.
        |
hypotheses_808080_total.out                          |

|-------------------------------------------------------------|
        | SPKR | # Snt # Wrd | Corr    Sub    Del
Ins    Err  S.Err |
```

```
|------+-------------+-----------------------------------|
       | bg   |   77    253 |  253       0       0
0      0       0 |

|------+-------------+-----------------------------------|
       | bk   |   77    253 |  252       1       0
0      1       1 |

|------+-------------+-----------------------------------|
       ...

|=============================================================|
       | Sum  |12547  41220 |41061    138     21
104    263    241 |

|=============================================================|
       | Mean | 77.0  252.9 |251.9    0.8    0.1
0.6    1.6    1.5 |
       | S.D. |  0.2    1.1 |  2.9    2.5    0.4
1.7    3.5    3.0 |
       |Median| 77.0  253.0 |253.0    0.0    0.0
0.0    1.0    1.0 |

`-------------------------------------------------------------'
```

EXPERIMENTAL DESIGN: STATISTICAL SIGNIFICANCE

Why is this important?

Click here if you want to learn more about how to measure statistical significance.