ASR Evaluation

Julia Hirschberg

CS 4706

Outline

- Intrinsic Methods
  - Transcription Accuracy
    - Word Error Rate
    - Automatic methods, toolkits
    - Limitations
  - Concept Accuracy
    - Limitations
- Extrinsic Methods

Evaluation

- How to evaluate the ‘goodness’ of a word string output by a speech recognizer?
- Terms:
  -

Evaluation

- How to evaluate the ‘goodness’ of a word string output by a speech recognizer?
- Terms:
  - ASR hypothesis: ASR output
  - Reference transcription: ground truth – what was actually said

Transcription Accuracy

- Word Error Rate (WER)
  - Minimum Edit Distance: Distance in words between the ASR hypothesis and the reference transcription
    - Edit Distance: = (Substitutions+Insertions+Deletions)/N
    - For ASR, usually all weighted equally but different weights can be used to minimize difference types of errors
  - WER = Edit Distance * 100

WER Calculation

- Word Error Rate =

100 (Insertions+Substitutions + Deletions)

Total Word in Correct Transcript

Alignment example:

REF: portable **** PHONE UPSTAIRS last night so

HYP: portable FORM OF STORES last night so

Eval I S S

WER = 100 (1+2+0)/6 = 50%

- Word Error Rate =

100 (Insertions+Substitutions + Deletions)

Total Word in Correct Transcript

Alignment example:

REF:	portable	****	phone	upstairs	last	night	so	***		
HYP:	preferable	form	of	stores	next	light	so	far		
Eval		S	I	S	S	S	S		I	

WER = 100 (1+5+1)/6 = 117%

NIST sctk-1.3 scoring softare: Computing WER with sclite

- http://www.nist.gov/speech/tools/
- Sclite aligns a hypothesized text (HYP) (from the recognizer) with a correct or reference text (REF) (human transcribed)

```
id: (2347-b-013)
Scores: (#C #S #D #I) 9 3 1 2
REF: was an engineer SO I i was always with **** **** MEN UM
   and they
HYP: was an engineer ** AND i was always with THEM THEY ALL
   THAT and they
Eval:         D S           I  I  S  S
```

Sclite output for error analysis

```
CONFUSION PAIRS                Total                            (972)
                             With >=   1 occurances             (972)
   1:   6 ->  (%hesitation) ==> on
   2:   6 -> the ==> that
   3:   5 -> but ==> that
   4:   4 -> a ==> the
   5:   4 -> four ==> for
   6:   4 -> in ==> and
   7:   4 -> there ==> that
   8:   3 -> (%hesitation) ==> and
   9:   3 -> (%hesitation) ==> the
  10:   3 -> (a-) ==> i
  11:   3 -> and ==> i
  12:   3 -> and ==> in
  13:   3 -> are ==> there
  14:   3 -> as ==> is
  15:   3 -> have ==> that
  16:   3 -> is ==> this
```

Sclite output for error analysis

```
17:  3 -> it ==> that
18:  3 -> mouse ==> most
19:  3 -> was ==> is
20:  3 -> was ==> this
21:  3 -> you ==> we
22:  2 -> (%hesitation) ==> it
23:  2 -> (%hesitation) ==> that
24:  2 -> (%hesitation) ==> to
25:  2 -> (%hesitation) ==> yeah
26:  2 -> a ==> all
27:  2 -> a ==> know
28:  2 -> a ==> you
29:  2 -> along ==> well
30:  2 -> and ==> it
31:  2 -> and ==> we
32:  2 -> and ==> you
33:  2 -> are ==> i
34:  2 -> are ==> were
```

Other Types of Error Analysis

- What speakers are most often misrecognized (Doddington ’98)
  - Sheep: speakers who are easily recognized
  - Goats: speakers who are really hard to recognize
  - Lambs: speakers who are easily impersonated
  - Wolves: speakers who are good at impersonating others

- What (context-dependent) phones are least well recognized?
  - Can we predict this?
- What words are most confusable (confusability matrix)?
  - Can we predict this?

Are there better metrics than WER?

- WER useful to compute transcription accuracy
- But should we be more concerned with meaning ("semantic error rate")?
  - Good idea, but hard to agree on approach
  - Applied mostly in spoken dialogue systems, where semantics desired is clear
  - What ASR applications will be different?
    - Speech-to-speech translation?
    - Medical dictation systems?

Concept Accuracy

- Spoken Dialogue Systems often based on recognition of Domain Concepts
- Input: I want to go to Boston from Baltimore on September 29.
- Goal: Maximize concept accuracy (total number of domain concepts in reference transcription of user input)

Concept	Value
Source City	Baltimore
Target City	Boston
Travel Date	Sept. 29

– CA Score: How many domain concepts were correctly recognized of total N mentioned in reference transcription

  Reference: I want to go from Boston to Baltimore on September 29

  Hypothesis: Go *from* Boston *to* Baltimore on December 29

  - 2 concepts correctly recognized/3 concepts in ref transcription * 100 = 66% Concept Accuracy

– What is the WER?

  - 3 Ins+2 Subst+0Del/11 * 100 = 45% WER (55% Word Accuracy)

Sentence Error Rate

- Percentage of sentences with at least one error
  - Transcription error
  - Concept error

Which Metric is Better?

- Transcription accuracy?
- Semantic accuracy?

Next Class

- Human speech perception