Parser Evaluation

L645 / B659

Dept. of Linguistics, Indiana University
Fall 2015
Some “easy” measures include:

- **accuracy**: how many sentences are correctly parsed?
- **coverage**: how many sentences can be parsed by the grammar?
- **n-best accuracy**: for how many sentences is the correct parse among the $n$ best parses?
PARSEVAL Measures

- PARSEVAL: Workshop in 1991 to decide how to evaluate parsers
- more “exact” measures, looking at single constituents
  - correct constituent: has the correct yield
- gold standard: correctly annotated data
PARSEVAL Measures (2)

- **precision**: number of correct constituents (yield) in parser output divided by number of constituents in the parser output

- **recall**: number of constituents from the gold standard (yield) that can be found in the parser output divided by the number of constituents in the gold standard

- **labeled precision**: percentage of correct constituents (yield + label) in parser output

- **labeled recall**: percentage of constituents from the gold standard (yield + label) that can be found in the parser output

- **F-score**: \( F_\beta = \frac{(\beta^2+1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \)
PARSEVAL Measures – Example

gold:

```
PX
   PX
   APPR
   an
   ART
   einem
   NN
   Samstag

PX
   ADVX
   NX
   vielleicht
   NX
   vielleicht
   oder
   NN
   Sonntag
```

parse:

```
PX
   APPR
   an
   ART
   einem
   NN
   Samstag

PX
   ADVX
   NX
   KON
   oder
   NN
   Sonntag
```
PARSEVAL Measures – Example

gold:

```
PX
 /   \
|     |
PX   ADVX
|     |     |
APPR NX ADV
|     |     |
an ART NX vielleicht
|     |     |
einem NX KON NX NN oder NN
|     |     |     |     |
Samstag Sonntag
```

parse:

```
PX
 /   \
|     |
PX   ADVX
|     |     |
APPR NX KON NX ADV
|     |     |     |
an ART NN oder NN vielleicht
|     |     |     |
einem Samstag Sonntag
```

precision: 3/6 = 0.5
recall: 3/7 = 0.42
F-score: $\frac{2 \times 0.5 \times 0.42}{0.5 + 0.42} = 0.457$
Other metrics

  - Obtain each word’s path to the root $S$
  - Compare to the gold standard, using edit distance

- Dependency evaluation (Lin 1995 and others):
  - Convert phrases to dependencies and evaluate percentage of correct dependencies
  - Does a better job of evaluating on “important” attachments and not penalizing parsers for propagating errors