| **Task** | **Evaluation Parameters** |
|---|---|
| **POS tagging:** | **Method:**<br>Precision and Recall for all POS tag classes<br>Weighted average of all precision and recall<br>Harmonic mean of Precision and Recall: FB1<br>**Reporting Figures:**<br>Overall FB1 measure<br>**Conditions and Assumptions:**<br>None<br>**Evaluation Token:** Word |
| **Chunking** | **Method:**<br>Precision and recall for all chunk labels.<br>Weighted Average and FB1 measure<br>**Reporting Figures:**<br>Overall FB1 measure<br>**Conditions and Assumptions:**<br>Chunker to be run with Gold Standard POS tags input<br>**Evaluation Token:** Chunk (boundary and label) |
| **Head Computation** | Accuracy computation taking chunk-heads as evaluation tokens<br>**Method:**<br>If the chunk head in GS matches a chunk head in output: its a Hit  if the chunk head in GS does not |

match the chunk head in output: Its a Miss

**Reporting Figures:**

hits / (hits +misses)

**Conditions and Assumptions:**

None

**Evaluation Token:** Chunk Heads

**Vibhakti Computation**

Accuracy computation taking contents of vibhakti field in chunks FS as evaluation tokens

**Method:**

If the vibhakti of chunk in GS matches the one in output: its a Hit if the vibhakti of chunk in GS does not match the one in output: Its a Miss

**Reporting Figures:**

hits / (hits +misses)

**Conditions and Assumptions:**

None

**Evaluation Token:** Vibhakti field in the chunk FS

**Morph**

Coverage and Accuracy computation over 5 categories of output

**Method:**

Comparing an FS with another: If all 8 fields match: Hit Else:Miss

Comparing a set of multiple FS with another set:

All Hits + no Misses: All correct

| | |
|---|---|
| All Hits + some Misses: | Mix Bag #1 |
| Some Hits + No Misses: | Mix Bag #2 |
| Some Hits + some Misses: | Mix Bag #3 |
| No Hits + All Misses: | All Wrong |

The Morph output for every word falls in one of above mentioned categories. And, the percentages of words falling into each of the categories is calculated.

**Reporting Figures:**

1. Numbers for all 5 categories as respective percentages/portions of morph output.
2. Coverage

**Conditions and Assumptions:**

Spelling Normalization

**Evaluation Token:** Feature Structure

## Glossary

**1. Hit:** If the systems' output for a token matches the expected output in the reference data, its a 'Hit'.

**2. Miss:** If the systems' output for a token does not match the expected output int he reference data, its a 'Miss'.

**3. Evaluation Token:** An evaluation token is the single smallest entity of output over which a '*Hit*' or a '*Miss*' is defined.

For example, while evaluating the POS tagger, the evaluation token would be the '*word*'. This would mean that, a word could be either correctly tagged or incorrectly tagged, but not partially correctly tagged.

Similarly, when evaluating the Chunker, the evaluation token would be the '*chunk*'. Meaning that a whole chunk(boundary and label) could either be correctly identified or incorrectly, but not partially correctly.

**4. Precision**: For any classification system, precision is defined as the (Number of correct predictions of the class /Number of total predictions for the class). For example, consider the outcome of a classification system

| ***Token:*** | ***Predicted class;*** | ***Reference class:*** |
|---|---|---|
| W1 | C1 | C1 |
| W2 | C2 | C2 |
| W3 | C3 | C4 |
| W4 | C4 | C3 |
| W5 | C1 | C2 |
| W6 | C1 | C2 |
| W7 | C2 | C2 |
| W8 | C3 | C3 |
| W9 | C4 | C1 |

The precision for every class would be

| ***Class*** | ***:*** | ***Number of correct predictions for class/ Number of predictions of the class*** |
|---|---|---|
| C1 | : | 1/3 |
| C2 | : | 2/2 |
| C3 | : | 1/2 |
| C4 | : | 0/2 |

Overall precision is the weighted average of precision across all classes.

**5. Recall**: For any classification system, Recall for any class is defined as the (Number of correct predictions of the class/ Total size of reference data for the class). Hence in the above example the recall figures would be

| ***Class*** | ***:*** | ***Number of correct predictions for class/ Size of the reference data for the class*** |
|---|---|---|
| C1 | : | 1/2 |
| C2 | : | 2/4 |
| C3 | : | 1/2 |
| C4 | : | 0/1 |

Overall precision is the weighted average of precision across all classes.

**6. FB1 measure:** FB1 measure is defined as the harmonic mean of precision and recall.

**7. Coverage** (or the extent of coverage)**:** Coverage of a system is defined as percent of input data for which the system produces an output.

Thus if a system produces no output for 1 out of 100 input tokens, the coverage is 99%.

**8. Accuracy:** Accuracy of a system is defined as the (Total number of correct outcomes / Total number of outcomes). In cases where the system does not produce an output for every input, the variations in the evaluation method include,

1. Reporting the coverage of the system separately.
2. Considering the failure to generate an output as a miss