Statistics

Untagged Corpus

test: 250,069
train: 20,187,105
valid: 250,007

Vocab size:
test: 17,849
train: 84,266
valid: 18,461

Tagged Corpus

test: 250,069
train: 20,187,105
valid: 250,007

Vocab size:
test: 17,469
train: 83,783
valid: 18,113

Four classes of words replaced by tags

<amt>
test: 3,832
train: 221,879
valid: 3,791

<sw>
test: 76,381
train: 6,116,217
valid: 76188

<dec>
test: 29

train: 1,786
valid: 37

<year>
test: 1,536
train: 130,915
valid: 1,418