Roger Bock
9/29/04

Assignment 2
SIMS 290-2: Applied Natural Language Processing

Data

I chose to work with the Penn Treebank dataset since it contains a reasonably large collection of tagged sentences. I decided to focus on the verb “say” since it occurs frequently throughout the corpus. I extracted all the tagged sentences which contained “say”, “said”, “saying”, or “says”. This gave me 493 sentences to work with.

Chunking

In this section I will list the different rules that I added to the existing chunking code. For each rule, I will give a description of the rule, the code containing the regular expression for the rule, and a before and after example.

Creating the rules by hand made me appreciate why systems that use statistical learning can be better than rules-based expert systems. Every time I fixed one problem, I felt like I either created or discovered another one.

Rule: Second NP chunking pass that turns <NP><PP> into <NP>
Code:

```
    rule = ChunkRule(r'<NP><PP>', 'Merge <NP><PP> into <NP>')
```

Before:

```
 (NP: <The/DT> <recent/JJ> <explosion/NN>)
 (PP: <of/IN> (NP: <country/NN> <funds/NNS>))
```

After:

```
 (NP:
   (NP: <The/DT> <recent/JJ> <explosion/NN>)
   (PP: <of/IN> (NP: <country/NN> <funds/NNS>)))
```

Rule: Include surrounding quotation marks in NP chunks

Code:

```
    quotationMarksRule =
        ChunkRule(r'(<DT|PRP.>?<RB>?)?<``><JJ|CD>*(<JJ|CD><,>)*(<NN.*>)+<
    \'\'>', 'Chunk quoted NPs')
```

Before:

```
 <the/DT>
 <``/``>
 (NP: <closed-end/JJ> <fund/NN> <mania/NN>)
 <"/">
```

After:

```
   (NP:
     <the/DT>
     <``/``>
     <closed-end/JJ>
     <fund/NN>
     <mania/NN>
     <"/">)
```

Rule: Allow possessives to join two NP chunks

Code:

```
    posRule1 = ChunkRule(r'<POS>', 'Chunk possessives')
    posRule2 =
        MergeRule(r'<.*>', r'<POS>', 'Include possessives with preceding NPs')
    posRule3 =
        MergeRule(r'<POS>', r'<.*>', 'Include possessives with proceeding NPs')
```

Before:

```
 (NP: <Newgate/NNP>)
 <'s/POS>
 (NP: <Mr./NNP> <Foot/NNP>)
```

After:

```
 (NP: <Newgate/NNP> <'s/POS> <Mr./NNP> <Foot/NNP>)
```

Rule: Allow adjectives after a verb
Code:

```
    rule =
        ChunkRule(r'<MD>?(<VB.*>)+(<RP>)?(<JJ>)*(<JJ>|<NP>|<PP>)+',
    'Chunk VPs')
```

Before:

```
 (NP: <People/NNS>)
 <have/VBP>
 <grown/VBN>
 <tired/JJ>
 (PP: <of/IN> (NP: <these/DT> <ads/NNS>))
```

After:

```
 (NP: <People/NNS>)
 (VP:
   <have/VBP>
   <grown/VBN>
   <tired/JJ>
   (PP: <of/IN> (NP: <these/DT> <ads/NNS>)))
```

Rule: Allow verb phrases with verbs in their infinitive form
Code:

```
    rule =
        ChunkRule(r'<MD>?(<VB.*><TO>)?(<VB.*>)+(<RP>)?(<JJ>)?(<JJ>|<NP>|
    <PP>)+', 'Chunk VPs')
```

Before:

```
 <your/PRP$>
 (NP: <TV/NN> <ad/NN>)
 <needs/VBZ>
 <to/TO>
 <be/VB>
 <bold/JJ>
```

After:

```
  <your/PRP$>
 (NP: <TV/NN> <ad/NN>)
 (VP: <needs/VBZ> <to/TO> <be/VB> <bold/JJ>)
```

Rule: Allow possessive pronouns to start noun phrases
Code:

```
    rule =
        ChunkRule(r'(<DT|PRP.>?<RB>?)?<JJ|CD>*(<JJ|CD><,>)*(<NN.*>)+',
    'Chunk NPs')
```

Before:

```
 <your/PRP$>
 (NP: <TV/NN> <ad/NN>)
```

After:

```
 (NP: <your/PRP$> <TV/NN> <ad/NN>)
```

Verb Argument Structure

I collected frequency information for the word and part of speech preceding the verb, the word and part of speech following the verb, and the phrase type following the verb. For the latter, when no phrase type was available, I reported the tag of the word following the verb. The results are reported at the end of this section, where I show the top ten entries in each frequency distribution.

For the previous words, it is not surprising that quotation marks and commas top the list. These correspond to sentences of the form:

“The weather is nice today”, said the person.

“The weather is nice today” said the person.

Both the previous words list and the previous tags list show that personal pronouns often occur before the verb. Again, this is not surprising, since I expected to find sentences of the form:

He said the weather is nice today.

For the next words and tags, I found that personal pronouns, proper nouns, and determiners top the list. These are again unsurprising, and correspond to sentences like:

He said it is likely to rain. (PRP)

“It’s true”, said John. (PPN)

He says the company’s prospects are dim. (DT)

Finally, analyzing the type of phrase found as the object of the verb shows that noun phrases top the list. These correspond to sentences like:

“It’s true”, said the company’s spokesperson.

```
Previous words:
              '' 65
               , 65
              he 26
         company 18
            also 14
              He 11
             she  9
           Corp.  8
              to  7
```

```
Previous tags:
             NNP 144
              '' 65
               , 65
             PRP 64
              NN 62
             NNS 48
              RB 20
              TO  7
              CC  6
Next words:
              it 70
               . 69
             the 61
               , 42
            that 39
             Mr. 16
            they 15
              he 11
             its 10
Next tags:
             PRP 107
             NNP 98
              DT 73
               . 69
              IN 47
               , 42
              NN 12
            PRP$ 12
              JJ  9
Next phrase:
              NP 211
             PRP 107
               . 69
               , 42
              PP 37
              IN 10
              EX  3
               :  2
            PRP$  2
```