

Kadyrov_Daniel_Assignment6

November 15, 2020

Daniel Kadyrov

Stevens ID: 10455680

CS557 - Natural Language Processing

Group 32 - Daniel Kadyrov

1 Part 1

Show your responses to the Your Turn practice of BKL Ch5 on pages 180. Try a variant of this exercise with the use of the two words “in” and “on”, as in “The book is not on the table, it is in my backpack”. Also tag the following sentence “Book the cooks who cook the books.” (a tagger based on the WSJ Corpus might be appropriate).

```
[6]: import nltk

text = "I like to swim in the swimming pool"
text = nltk.word_tokenize(text)
nltk.pos_tag(text)
```

```
[6]: [('I', 'PRP'),
      ('like', 'VBP'),
      ('to', 'TO'),
      ('swim', 'VB'),
      ('in', 'IN'),
      ('the', 'DT'),
      ('swimming', 'NN'),
      ('pool', 'NN')]
```

```
[7]: text = "The book is not on the table, it is in my backpack"
text = nltk.word_tokenize(text)
nltk.pos_tag(text)
```

```
[7]: [('The', 'DT'),
      ('book', 'NN'),
      ('is', 'VBZ'),
      ('not', 'RB'),
      ('on', 'IN'),
```

```
( 'the', 'DT'),
( 'table', 'NN'),
( ',', ', '),
( 'it', 'PRP'),
( 'is', 'VBZ'),
( 'in', 'IN'),
( 'my', 'PRP$'),
( 'backpack', 'NN')]
```

```
[8]: text = "Book the cooks who cook the books"
text = nltk.word_tokenize(text)
nltk.pos_tag(text)
```

```
[8]: [('Book', 'IN'),
      ('the', 'DT'),
      ('cooks', 'NNS'),
      ('who', 'WP'),
      ('cook', 'VBP'),
      ('the', 'DT'),
      ('books', 'NNS')]
```

2 Part 2

Respond to BKL Exercise 5.1 on p215. Tag the following sentence “Book the cooks who cook the books.” (a tagger based on the WSJ Corpus might be appropriate) and several other sentences you can think of with ambiguous words, as suggested by this exercise.

2.1 Exercise 5.1

Search the Web for “spoof newspaper headlines,” to find such gems as: British Left Waffles on Falkland Islands, and Juvenile Court to Try Shooting Defendant. Manually tag these headlines to see whether knowledge of the part-of-speech tags removes the ambiguity.

```
[9]: headline = "Juvenile Court to Try Shooting Defendant"
headline = nltk.word_tokenize(headline)
nltk.pos_tag(headline)
```

```
[9]: [('Juvenile', 'NNP'),
      ('Court', 'NNP'),
      ('to', 'TO'),
      ('Try', 'VB'),
      ('Shooting', 'NNP'),
      ('Defendant', 'NNP')]
```

```
[10]: text = "Book the cooks who cook the books"
text = nltk.word_tokenize(text)
nltk.pos_tag(text)
```

```
[10]: [('Book', 'IN'),
      ('the', 'DT'),
      ('cooks', 'NNS'),
      ('who', 'WP'),
      ('cook', 'VBP'),
      ('the', 'DT'),
      ('books', 'NNS')]
```

```
[11]: text = "Panda eats shoots, and leaves"
      text = nltk.word_tokenize(text)
      nltk.pos_tag(text)
```

```
[11]: [('Panda', 'NNP'),
      ('eats', 'VBZ'),
      ('shoots', 'NNS'),
      (',', ','),
      ('and', 'CC'),
      ('leaves', 'NNS')]
```

```
[12]: text = "Panda eats shoots and leaves"
      text = nltk.word_tokenize(text)
      nltk.pos_tag(text)
```

```
[12]: [('Panda', 'NNP'),
      ('eats', 'VBZ'),
      ('shoots', 'NNS'),
      ('and', 'CC'),
      ('leaves', 'NNS')]
```

```
[13]: text = "Include your children when baking cookies"
      text = nltk.word_tokenize(text)
      nltk.pos_tag(text)
```

```
[13]: [('Include', 'NNP'),
      ('your', 'PRP$'),
      ('children', 'NNS'),
      ('when', 'WRB'),
      ('baking', 'VBG'),
      ('cookies', 'NNS')]
```

3 Part 3

Respond to Exercise 12.10 from J&M 2nd Ch12 for at least one Wh- terminal. The lexicon and grammar in Figures 10.2 and 10.3 should be enough. If not indicate any other terminals and rules necessary. Obviously the Wh- you chose needs to be a terminal. You are welcome to search the WWW for help here; be sure to show references to the sites you use.

3.1 Exercise 12.10

Page 397 discussed the need for a Wh-NP constituent. The simplest Wh-NP is one of the Wh-pronouns (who, whom, whose, which). The Wh-words what and which can be determiners: which four will you have?, what credit do you have with the Duke? Write rules for the different types of Wh-NPs.

WH-NP -> Wh-Pro

WH-NP -> Wh-Det Nominal

Wh-Pro -> who, whom, whose, which

Wh-Det -> what, which

4 Part 4

Chapter 10 of J&M 3rd and Chapter 5 of BKL demonstrate several taggers. Make a list of them and any others you can find with a brief statement as to what they do, what is distinctive about them, and what is like the other taggers in the chapter. You can show them with a table of at least 4 columns: ID, Description, Distinctive Feature(s), and “Similar To”. The Description should include words used in both J&M 2nd and 3rd and BKL, such as “bidirectional MEMM-style”, “regular expression tagger”, “lookup tagger”, etc. to distinguish the taggers.

ID	Description	Distinctive Features	Similar To
Default Tagger	POS Tagger	POS, NLTK	POS
POS Tagger	Part of Speech. Processes a sequence of words and attaches a part of speech tag to each word	Speech Tag , NLTK	POS
Regular Expression	Assigns tags to tokens based on matching patterns	POS, NLTK	POS
Lookup Tagger	Initially uses a lookup table to assign tokens that the tagger identified as None. Then uses backoff, where one tagger acts as a parameter to another.	Regex, NLTK	POS
Unigram Tagger	Each token gets assign a tag that is most likely for that particular token	POS, isolated word	Lookup Tagger
N-Gram Tagger	Uses n-1 preceding words to give context to the current word		Unigram Tagger
Brill Tagger	Guesses the tag of each word, then goes back to fix the mistake	Supervised Learning	
MEMM	Maximum Entropy Markov Model. Uses logistic regression to classify words based on the words in its succession		

5 Part 5

Tag the longest “sentence” you found from Assignment 2. Compare the results for at least two different taggers. Briefly discuss what applying embeddings to this might produce.

```
[38]: # From Assignment 2
from nltk.corpus import inaugural
from nltk.tokenize import sent_tokenize

sents = sent_tokenize(inaugural.raw())
max_sent = max(sents, key=len)
max_sent_toks = nltk.word_tokenize(max_sent)

max_sent
```

```
[38]: 'On this subject it might become me better to be silent or to speak with
diffidence; but as something may be expected, the occasion, I hope, will be
admitted as an apology if I venture to say that if a preference, upon principle,
of a free republican government, formed upon long and serious reflection, after
a diligent and impartial inquiry after truth; if an attachment to the
Constitution of the United States, and a conscientious determination to support
it until it shall be altered by the judgments and wishes of the people,
expressed in the mode prescribed in it; if a respectful attention to the
constitutions of the individual States and a constant caution and delicacy
toward the State governments; if an equal and impartial regard to the rights,
interest, honor, and happiness of all the States in the Union, without
preference or regard to a northern or southern, an eastern or western, position,
their various political opinions on unessential points or their personal
attachments; if a love of virtuous men of all parties and denominations; if a
love of science and letters and a wish to patronize every rational effort to
encourage schools, colleges, universities, academies, and every institution for
propagating knowledge, virtue, and religion among all classes of the people, not
only for their benign influence on the happiness of life in all its stages and
classes, and of society in all its forms, but as the only means of preserving
our Constitution from its natural enemies, the spirit of sophistry, the spirit
of party, the spirit of intrigue, the profligacy of corruption, and the
pestilence of foreign influence, which is the angel of destruction to elective
governments; if a love of equal laws, of justice, and humanity in the interior
administration; if an inclination to improve agriculture, commerce, and
manufacturers for necessity, convenience, and defense; if a spirit of equity and
humanity toward the aboriginal nations of America, and a disposition to
meliorate their condition by inclining them to be more friendly to us, and our
citizens to be more friendly to them; if an inflexible determination to maintain
peace and inviolable faith with all nations, and that system of neutrality and
impartiality among the belligerent powers of Europe which has been adopted by
this Government and so solemnly sanctioned by both Houses of Congress and
applauded by the legislatures of the States and the public opinion, until it
shall be otherwise ordained by Congress; if a personal esteem for the French
```

nation, formed in a residence of seven years chiefly among them, and a sincere desire to preserve the friendship which has been so much for the honor and interest of both nations; if, while the conscious honor and integrity of the people of America and the internal sentiment of their own power and energies must be preserved, an earnest endeavor to investigate every just cause and remove every colorable pretense of complaint; if an intention to pursue by amicable negotiation a reparation for the injuries that have been committed on the commerce of our fellow-citizens by whatever nation, and if success can not be obtained, to lay the facts before the Legislature, that they may consider what further measures the honor and interest of the Government and its constituents demand; if a resolution to do justice as far as may depend upon me, at all times and to all nations, and maintain peace, friendship, and benevolence with all the world; if an unshaken confidence in the honor, spirit, and resources of the American people, on which I have so often hazarded my all and never been deceived; if elevated ideas of the high destinies of this country and of my own duties toward it, founded on a knowledge of the moral principles and intellectual improvements of the people deeply engraven on my mind in early life, and not obscured but exalted by experience and age; and, with humble reverence, I feel it to be my duty to add, if a veneration for the religion of a people who profess and call themselves Christians, and a fixed resolution to consider a decent respect for Christianity among the best recommendations for the public service, can enable me in any degree to comply with your wishes, it shall be my strenuous endeavor that this sagacious injunction of the two Houses shall not be without effect.'

```
[41]: from nltk.corpus import brown

# Regular Expression Tagger

patterns = [
    (r'.*ing$', 'VBG'), # gerunds
    (r'.*ed$', 'VBD'), # simple past
    (r'.*es$', 'VBZ'), # 3rd singular present
    (r'.*ould$', 'MD'), # modals
    (r'.*\'s$', 'NN$'), # possessive nouns
    (r'.*s$', 'NNS'), # plural nouns
    (r'^-?[0-9]+(.[0-9]+)?$', 'CD'), # cardinal numbers
    (r'.*', 'NN') # nouns (default)
]

regex_tagger = nltk.RegexpTagger(patterns)
regex_tagger.tag(max_sent_toke)[:20]
```

```
[41]: [('On', 'NN'),
      ('this', 'NNS'),
      ('subject', 'NN'),
      ('it', 'NN'),
```

```

('might', 'NN'),
('become', 'NN'),
('me', 'NN'),
('better', 'NN'),
('to', 'NN'),
('be', 'NN'),
('silent', 'NN'),
('or', 'NN'),
('to', 'NN'),
('speak', 'NN'),
('with', 'NN'),
('diffidence', 'NN'),
(';', 'NN'),
('but', 'NN'),
('as', 'NNS'),
('something', 'VBG')]

```

[42]: *# Lookup Tagger*

```

fd = nltk.FreqDist(brown.words(categories="news"))
cfd = nltk.ConditionalFreqDist(brown.tagged_words(categories="news"))

most_freq_words = list(fd.keys())

likely_tags = dict((word, cfd[word].max()) for word in most_freq_words)

baseline_tagger = nltk.UnigramTagger(model=likely_tags, backoff=nltk.
↳DefaultTagger("NN"))

baseline_tagger.tag(max_sent_tokenize[:20])

```

[42]:

```

[('On', 'IN'),
 ('this', 'DT'),
 ('subject', 'NN'),
 ('it', 'PPS'),
 ('might', 'MD'),
 ('become', 'VBN'),
 ('me', 'PPO'),
 ('better', 'JJR'),
 ('to', 'TO'),
 ('be', 'BE'),
 ('silent', 'JJ'),
 ('or', 'CC'),
 ('to', 'TO'),
 ('speak', 'VB'),
 ('with', 'IN'),
 ('diffidence', 'NN'),

```

```
(';', ' '),  
( 'but', 'CC'),  
( 'as', 'CS'),  
( 'something', 'PN')]
```

The Lookup tagger tested performed better than the Regular Expression tagger at tagging the longest sentence of the inaugural address. As the regex patterns increase, this might be fixed but that requires predicting and identifying every possible pattern versus the lookup tagger learns as its baseline tagger and backoff memory increase.