CSC485 - A3

Junyi Sun

December 2016

Please use Python3.5 to run my programs. Thank you

1 Word Sense Disambiguation

1. The fine-grainedness of WordNet gives a lot of details and definitions of the word. The fine-grainedness property provides a lot more choices of senses for a normal word, which has pros and cons, the pros are: for some non-often-used words in English, we can still find multiple definitions for them in WordNet, enable us to find the most relevant word-sense among all senses; cons: because the WordNet is too detailed, it contains a lot of rare senses of certain words, which often leads to ambiguity because when we use lesk algorithms, we treat the rarely-used senses as important as the most-used senses, sometimes the rarely-used senses of one word accidentally overlap with the rarely-used senses of another word in the same sentence, then it will have the most count and cause mistake. For example, one rare sense of 'addition' has something to do with 'suburban area' and one rare sense of 'greens' is the street name. Therefore, these two rare senses overlap the most and make this rare sense of 'addition' the most count, which is a mistake.

2. I would rank all senses of a word from most frequently used to rarely used, and only count the top most frequently appeared senses, delete those rarely used ones. I would delete the 'street name' sense of 'greens' and 'suburban area laid out in street' sense of 'additions' because these usages are so rare and leads to the accident overlap between these two words, make wrong suggestion for the sense of 'addition'.

3. By adding examples of the usage increases the priority the top ranked sense, if the top rank wos originally the correct one, it will produce the correct answer, however, if the top rank was wrong in the first place, adding examples won't fix the mistake but make it worse. For example, before adding words from example usage to the word bag, the count of avocado sense2 (tropical American tree) is 4, which is the top rank but wrong. After adding example words, the count of avocado sense2 becomes 5, which gets worse than before.

4. Yes I used the NLTK stopword list, and I added a few more words to it such as 'thing' and 'something' because those words are meaningless towards finding the most correct sense.

2 Hyperonym Relations

In order to count the number of occurrence of 4 cases for the nytbig, the program processes all these sentences and for each suggestions that fit into certain category, I print the category number into a new output file, and using online counting tools to count the number of occurance of each number in that file. I get the following result for nytbig:

Case '1' Count: 10054
Case '2' Count: 593
Case '3' Count: 38331
Case '4' Count: 10081
Total 59059

Case	High Confidence	Medium	Low Confidence
1	261	3116	6677
2	15	165	413
3	843	13415	24073
4	181	2315	7585

Please see my evaluation of suggestions in the attached output file for case 3 and 4

2.1 Success

1. when the sentence strictly follow the given pattern, the program will return the result correctly
   ex)Congress limits deductions for gifts of appreciated assets , such as stock , to 30 percent of the donor 's adjusted gross income .
   HYPONYM(stock, asset)

2. the program correctly identifies the compound noun words most of the time, distinguish them with 'det + noun' phrases, these determinants will not appear in the returned suggestions.
   ex) A version of the rule proposed a year and a half ago would have required lenders to offer a single price for the various costs of processing the mortgage – including such expenses as appraisals and title insurance – or to stick much more closely to initial estimates of those costs . HYPONYM(appraisal, expense) HYPONYM(title insurance, expense)

3. the program includes both direct hyperonyms and super hyperonyms of the words, becuase often when a compound word appears, the last noun is the super hyperonym.
   ex) HYPONYM(Cardinal Bernard F. Law, church official)
   HYPONYM(Cardinal Bernard F. Law, official)

HYPONYM(Google, search engine)
HYPONYM(Google, engine)

4. the program correctly removes all the meaningless adverbs that are included in the parsed tree noun phrase
ex) Over all , Asia , especially China , was responsible for the greatest increase in the purchase of United States products .
(NP Asia/NP) ,/, (NP especially/RB China/NP)
HYPONYM(China, Asia)
ex) While riding the rails does offer the freedom to phone , dropped calls are a big fear for many passengers , especially salespeople doing business .
(NP many/JJ passengers/NNS) ,/, (NP especially/RB salespeople/NNS)
HYPONYM(salesperson, passenger)

5. the program correctly process names even though these names are not in the WordNet
ex) The charges and countercharges flew one day after Dr. Dean 's campaign began showing a television commercial that criticized Mr. Gephardt and other candidates for supporting the Iraq conflict , part of a new aggressive tactic .
HYPONYM(Mr. Gephardt, candidate)

2.2   Shortcomings

1. When a supplemental phrase appears before the 'such as' identifier, although the last word is a noun , it is not the hyperonym of the group. The following example illustrate this: 'people' should be the correct hyperonym of 'force operative' and 'pilot', but instead, the last noun before 'such as' gets returned.
ex) ther firms under investigation , including One Group .
(NP investigation/NN) ,/, including/VBG (NP One/NP Group/NP)
HYPONYM(One Group, investigation)

2. Numbers are not identified and often included in the result suggestions. However, the optimal result should not include numbers in the noun
ex) HYPONYM(family member, 400 people)
HYPONYM(graphic artist, 60 designer)

3. When there are multiple noun phrases in parallel, the program cannot identify which noun phrase should be included in the suggestion and which shouldn't
ex)Upon the ribbon of 2,982 victims ' names , according to the development corporation , etched in stone about the memorial 's two reflecting pools , the police officers , firefighters and other rescue workers will be designated with individual shields .
HYPONYM(pool, rescue worker)

HYPONYM(police officer, rescue worker)
HYPONYM(firefighter, rescue worker)

4. Some sentences are not actually follow the given pattern, but the parsed structure are following the pattern, lead to wrong suggestions
   ex) He has worked in developing countries for 35 years , including a stint as the Ford Foundation 's representative in India .
   (NP 35/CD years/NNS) ,/, including/VBG (NP a/DT stint/NN) HYPONYM(stint, 35 year)

5. when there are descriptive clauses attached to the main NP, the noun in the clause will be mistakenly count as the hyperonym because it is closer to the identifier words(including, especially etc)
   ex) the list also includes several national organizations that are mainstays of communal life for many American Muslims , including the Muslim Student Association , the Islamic Society of North America and the Islamic Circle of North America .
   HYPONYM(Muslim Student Association, Muslims)

6. In family names, the determinant 'the' cannot be deleted, but the program automatically ignore such case. ex) Along with such prominent real estate families as the Rudins , the Dursts and the Tishmans , the Fishers helped fashion the Manhattan skyline , erecting some of New York 's most recognizable corporate towers .
   HYPONYM(Rudins, family)
   HYPONYM(Dursts, family)
   HYPONYM(Tishmans, family)