CS372 Homework #3

20160632 Janghyun Cho

Our goal is to find homographs and heteronyms in a sentence. First step was to set definition of homographs and heteronyms. I set them up from their definitions, given problem description, and ranking guidelines. Homograph is one of words in sentence that has same spelling but has different meaning from other word in same sentence. Heteronym is a word that has same spelling, just like homograph, but it must have different pronunciation. Thing here is that heteronyms can exist alone in sentence. From priority rule 2, wind + wind is higher than wind + tear that has only heteronyms but no homographs, I concluded that homographs must have its pair that has different meaning but heteronyms can go alone if only it has possibility of being pronounced differently.

For the goal, I'll use brown corpus for getting sentences, and use Wordnet with Wordnet Lemmatizer for figuring out homographs. Lastly, cmudict will be used for finding heteronyms. Based on these, my idea is like following. For given sentences, I tokenize the text and lemmatize them and save its original form. What we aim to find is homographs and heteronyms from nouns, verbs, adjectives, and adverbs. Homographs are words that have same lemmatized original form and have multiple meanings in synset along with different position tags. Heteronyms are words that have multiple pronunciations in cmudict obviously with multiple meanings. I manually removed set of verbs with no specific meanings such as be-verbs, do, and have. Next, give points according to ranking priority 1, 2, and 3. If they have equal points in all three cases, shorter sentence comes first. Last part is printing. First for 30 sentences with annotations and its category in brown corpus. Then for 30 sentences with given ranking rule. Lastly, for the heteronyms in the sentence, assign nth pronunciation method in cmudict depending on what parts-of-speech it is: noun, verb, adjective, or adverb.

Following are result I got by above implementation and rules. I got two random sentences from initial sentences and ranked sentences. 14th and 26th sentence from initial sentences and 12th and 25th sentence from ranked sentences. *14. Fortunately it spared us from the usual (Y UW1 ZH AH0 W AH0 L) spate of silly resolutions which in the past have made Georgia look like anything but `` the empire state of the South '' .* There is no words that occurs twice and found one heteronym: usual. Usual has two meanings and two pronunciations Y UW1 ZH AH0 W AH0 L and Y UW1 ZH UW0 AH0 L and usual in sentence is short usual. For 26th sentence, *26. Look to Coosa Valley for industrial progress (P R AA1 G R EH2 S).* There are no homographs and has a heteronym, progress that can have three pronunciations: P R AA1 G R EH2 S, P R AH0 G R EH1 S, and P R OW0 G R EH1 S. Given first pronunciation looks fairly good enough. Either first or third would be decent answer. For ranked sentences, 12th sentence is *12. Her creations in fashion are from many designers because she doesn't want (W AA1 N T) a complete wardrobe from any one designer any more than she wants (W AO1 N T S) `` all of her pictures by one painter (P EY1 N T ER0) '' .* There are two homographs designer, want, designer, and want. Designers can be interpreted by multiple meanings, one is a person designer and other is adjective meaning of designer. Word want also has multiple meanings. First is popular known meaning of want, and other is meaning of need. First want, (W AA1 N T) can be thought as having meaning need and second want, (W AO1 N T S) can be though as popular meaning of want. 25th sentence is like following. *25. Most children love the animated (AE1 N AH0 M EY2 T IH0 D) puppet (P AH1 P AH0 T) faces (F EY1 S IH0 Z) and their flexible bodies , and they prefer (P R AH0 F ER1) to see them as though the puppets (P AH1 P AH0 T S) were in action , rather (R AE1 DH ER0) than put away in boxes (B AA1 K S AH0 Z) .* Similar with above, words have duplicate pronunciations and meanings however there's lack of space to write for all of them. Just for the word puppet, which is a homograph and heteronym, can be thought as a doll or negative meaning of puppet. Also it has two pronunciations : P AH1 P AH0 T and P AH1 P IH0 T. Similar cases work for other words.

Following improvements can be made to improve it. The biggest and the part I can't figure how to be implemented throughout the three homework that we've done is that I can't figure out how to match specific meaning in given sentence to meaning in synset. Also synset has very specified meanings, so that it is segmented more than dictionary. Giving possibility to over finding homographs. However I couldn't get fully open source English dictionary so I implemented it with synset. In this homework additionally, definitions and parts-of-speech has to be matched to pronunciations. Currently I just manually found most matching pattern: nouns, adjectives, and adverbs get first pronunciation and verbs get second pronunciation. It would have been best to find a way to connect it to a group of words in synset.