[Overview]

In this homework, the final goal is to find out which person the selected pronouns are bound to. For example, in the sentence 'John asked Peter to help him', 'him' is bound to John, not Peter. Here, I used several approaches to solve the problem. First, for snippet-context, gender matching, count of the names, distance between the name and the pronoun, and the syntactic clues were mainly used. For the page-context, an additional factor, the total count of the names in the page was used. In order to wrap these up in one prediction, I used nltk MaxentClassifier with the extracted feature set including those features.

[Detailed Approach]

First, I will briefly explain about the dataset I used. GAP coreference dataset is a gender-balanced dataset containing coreference-labeled texts. Here, the only pronouns are 'he', 'his', 'him', 'she', 'her', and 'hers'. Pronouns such as 'I', 'my', 'and' are not contained in the dataset, and also reflexive pronouns such as 'himself', or 'herself' is not included. For page-context approach, url for wikipedia is given.

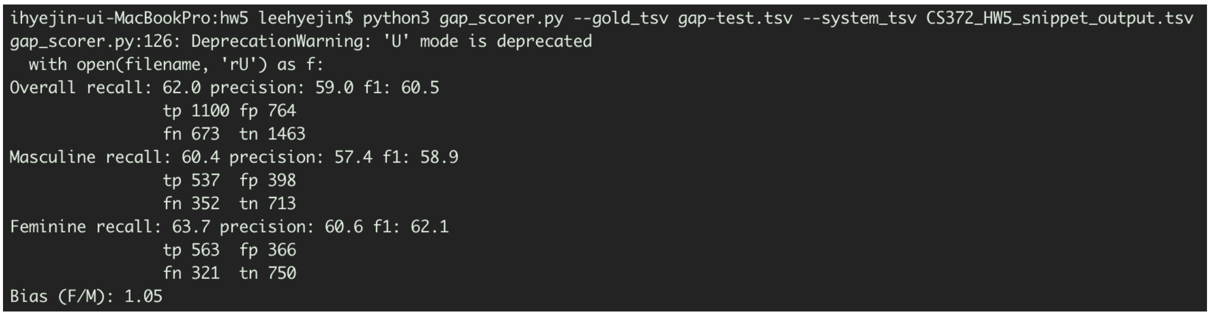
Here, I will explain the approaches one by one. First, gender matching was used. 'nltk.corpus.names' was used to find out male and female names. If the given name (A or B) is matched with only one of them, their gender was annotated as 'F' or 'M'. If the given name consists of both the last name and the first name, only the first name was used to annotate gender. Together with the gender annotation of the given names, given pronoun was annotated with gender, too. However, using this approach alone this made a lot of errors. The reason might be the characteristic of used dataset where gender matching doesn't really help. If the pronoun is male pronoun, then the candidate names were all males. Thus, in the final model, I deleted this factor from the feature set. Second, the occurrence of the names in given text was counted since it is highly likely for the pronoun to be bound to frequently used name. Third, distance between the name and the pronoun was used. (They are measured in word counts.) There is no one-to-one correspondence between the distance and the co-referencing relationship, but if two words are too far apart, or even too close, the possibility of coreference will be lower. Two words can be too far apart that the pronoun cannot be bound to the name, or if they are too close, one might be directly connected with the other one. For example, in the sentence 'John likes her', 'her' is not bound to 'John'. Instead, 'her' is directly connected to 'John', and is describing John's action. Since this dataset does not contain any reflexive pronouns, this approach is highly likely to work.

Since those approaches kind of worked, but I needed syntactical clues since the length of the sentences or phrases can grow infinitely long and then the approach using the word distance alone will not work anymore. Here, I used Stanford CoreNLPParser in order to tokenize, tag and parse the sentence and get the tree structure. One the tree structure of the sentence is ready, I measured two factors, tree distance and direct connection. Tree distance is the shortest path length between two words. This can be explained as an advanced approach, similar as the word distance. Since the tree distance shows the syntactic distance between two words, it is included in the feature set. Direct Connection measures another syntactic factor in Boolean, using the NP tags. As I shortly described above about the direct connection between the name and the pronoun, if they are directly connected, the pronoun is not likely to co-reference the name. Here, starting with the target pronoun, I went up to see the first NP or S tag it is in. Then, if the name is also under the tag, it was regarded as 'direct connection'. For example, in the sentence 'John said that Mary likes him.' 'Mary likes him' is parsed as S, and since that sentence is the first occurring S above 'him', and 'Mary' is under the S, Mary' and 'him' is regarded as direct connection. However, since 'John' is outside of the S, 'John' and 'him' is not regarded as direct connection.

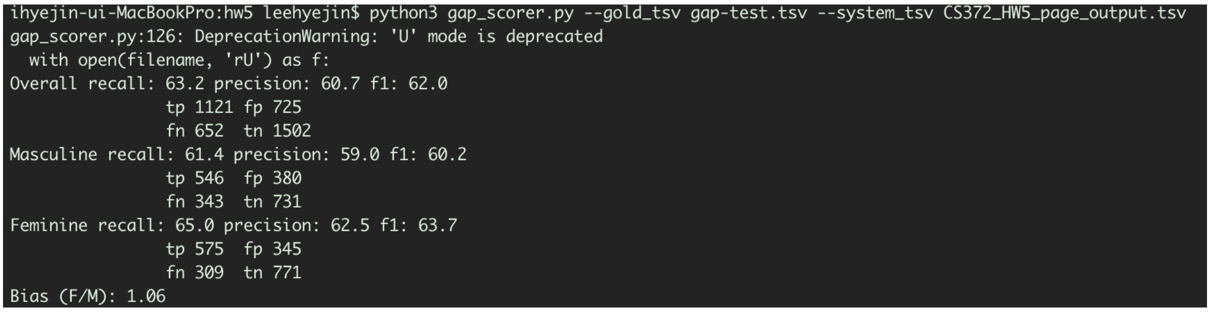
In addition, in the page-context approach, I used one more factor, the total count of the names in the page. Since the count in snippet context only showed a local information, the total count in the page was added in order to show a wider information of the name occurrence.

To wrap these information into one and predict the co-referencing relationship of given inputs, I used nltk MaxentClassifier, trained them with the 2000 pairs of extracted features in ‘gap-development.tsv’, and tested with the 2000 pairs of ‘gap-test.tsv’.

[Overall result]



```
ihyejin-ui-MacBookPro:hw5 leehyejin$ python3 gap_scorer.py --gold_tsv gap-test.tsv --system_tsv CS372_HW5_snippet_output.tsv
gap_scorer.py:126: DeprecationWarning: 'U' mode is deprecated
  with open(filename, 'rU') as f:
Overall recall: 62.0 precision: 59.0 f1: 60.5
                tp 1100 fp 764
                fn 673  tn 1463
Masculine recall: 60.4 precision: 57.4 f1: 58.9
                tp 537  fp 398
                fn 352  tn 713
Feminine recall: 63.7 precision: 60.6 f1: 62.1
                tp 563  fp 366
                fn 321  tn 750
Bias (F/M): 1.05
```



```
ihyejin-ui-MacBookPro:hw5 leehyejin$ python3 gap_scorer.py --gold_tsv gap-test.tsv --system_tsv CS372_HW5_page_output.tsv
gap_scorer.py:126: DeprecationWarning: 'U' mode is deprecated
  with open(filename, 'rU') as f:
Overall recall: 63.2 precision: 60.7 f1: 62.0
                tp 1121 fp 725
                fn 652  tn 1502
Masculine recall: 61.4 precision: 59.0 f1: 60.2
                tp 546  fp 380
                fn 343  tn 731
Feminine recall: 65.0 precision: 62.5 f1: 63.7
                tp 575  fp 345
                fn 309  tn 771
Bias (F/M): 1.06
```

[Limitations and Improvements]

Inside this approach, there are many ‘possible’ factors that can attribute to the co-referencing relationships. There are some limitations in this approach. First of all, since I used classifier to sum up all the possible factors, it is hard to exactly say that how much weight each factor attributes to the relationship. It also makes me hard to find out which factor doesn’t really work and which does really work. However, during the process of adding features, I could check that the accuracy gets higher from lower than 50 to about 60. Thus, I can say that these features represent at least some part of the relationship between the pronoun and each name.

The syntactic relationship I used was not really deep, too. As people can easily notice which word is the pronoun bound to, a deeper insight to the grammatical structure is needed. Here, finding the dependencies of the words can directly give the relationship between two phrases, so it will help finding the co-referencing relationship between the words.

Also, the characteristics of pronoun can be concerned in a deeper way. Subjective, objective and possessive pronouns will work differently in the sentence. By scanning the sentence structure, ‘him’ and ‘her’ can be classified into objective or possessive regard of their use in the structure. Thus, including a form of pronoun in the feature set will work.

Since this dataset only contains a small, limited set of coreference tasks, there were only a limited number of factors I had to consider. For the other tasks, there can be more approaches to try. First, gender and identity matching can help. For example, if the pronoun is ‘it’, then person name will not be the answer. Or if the pronoun is ‘he’, ‘Mary’ seems not to be the answer. The ‘pronoun’ can be more considered, too. Reflexive pronouns, such as ‘himself’, which are not considered in handling this dataset, can be considered, too.