이수로 20160830

Coreference resolution is crucial in natural language understanding because resolving ambiguous pronouns prevent machines from clearly interpreting the full meaning of the text. Our model aims to resolve coreferences by using a combination of heuristics to classify whether a pronoun refers to a potential antecedent.

First, we observed training data to see if there were any potential heuristics we could exploit. Each of these heuristics were evaluated based on its significance in classifying the correct antecedent in the validation data. Each heuristic is further explained below along with its evaluation. Then we present the set of selected heuristics based on the evaluation and its weighting in the total scoring of each pair. The score of each pair is then used to classify whether each (antecedent, pronoun) pair is correct.

1. Token Distance: Token distance is calculated by finding the distance between the pronoun token and the antecedent token. Although initially the antecedent was expected to be closer to the pronoun, token distance results in very random numbers due to cases like 'John, who loves playing baseball, hockey, and soccer with Jim, loves his dogs'. Also, there are other cases like 'She likes me. Sandy told me', where the pronoun refers to a previous sentence but ends up closer to an antecedent in the following sentence. From **Table 1** and **Table 2**, we show that not only are the distances very sporadic but also the means show no significant difference. As a result, token distance was excluded from test-time heuristics.

2. Parse Tree Distance: Parse tree distance was an attempt to find a replacement for token distance. First, a parse tree is created by AllenNLP's Constituency Parser. Then, the distance from the NP that contains the antecedent to the NP that contains the pronoun in the parse tree is calculated. Compared to token distance, parse tree distance is clearly more representative of the relations between the antecedent and the pronoun since syntactic closeness is considered. By observing the means from **Table 1** and **Table 2**, it is evident that correct pronoun-antecedent pairs have a significantly lower parse tree distance than incorrect pronoun-antecedent pairs. Thus, parse tree distance is used in test-time heuristics.

3. Word Frequency: Another intuitive heuristic was word frequency. Since the main topic of discourse is most likely to be continuously referred to as a pronoun, we compared the number of mentions of each correct and incorrect antecedent in the given text. In **Table 1** and **Table 2**, the means differ slightly, but not by a significant amount. As a result, we use word frequency in test-time heuristics, but use a lower weight so that word frequency is not the main decision factor.

4. Antecedent Grammatical Argument: Similar to the rationale behind word frequency, we hypothesized that a (antecedent, pronoun) pair is more likely to be correct if the antecedent is the subject of a sentence, and less likely to be correct if the antecedent is the object of a sentence. Thus, in the implementation, a subject was encoded as 1, an object as -1, and neither as 0. The significantly different means from **Table 1** and **Table 2** somewhat support this hypothesis. However, we observe that although correct pronoun-antecedent pairs are more likely to have subject antecedents, the results for incorrect pronoun-antecedent pairs are inconclusive. As a result, we use 1 to encode a subject, and 0 to encode anything that is not a subject during test-time heuristics and weigh the encoding.

5. Grammatical Argument Agreement: Grammatical argument agreement for a (antecedent, pronoun) pair is 1 if they are both subjects or both objects, and 0 otherwise. We hypothesized that a pronoun would have the same grammatical argument as its antecedent (ex. 'I like John, Bob likes him too.'). However, from **Table 1** and **Table 2**, we observe no significant difference in grammatical argument between a correct pair and an incorrect pair. As a result, we omit this metric from test-time heuristics.

6. Is_Constrained: The Is_Constrained value is 1 when the (antecedent, pronoun) pair is very unlikely to be correct, and 0 otherwise. An (antecedent, pronoun) pair is thus constrained when: 1. The pair are in the same NP (ex. ′her dog Mary′), and 2. The pair are in the same sentence, the pair are both possessive or both non-possessive, and one of the pair is the subject of the sentence and the other is the object (ex. ′Sally yelled at her′, ′Jim′s friend called his dad′). The condition that the pair have to be both possessive or both non-possessive exists because there are many sentences like ′Jim walked his dog′, where the pronoun is possessive and the antecedent is non-possessive. From **Table 1** and **Table 2**, it is evident that the criteria above is well chosen; no correct pairs have been marked as constrained, while some incorrect pairs have. Thus this heuristic is used in test time to decrease the number of false positives.

	Token Distance	Parse Tree Distance	Word Frequency	Antecedent Grammatical Argument	Same Grammatical Argument	Is_Constrained
Mean	15.57	6.77	1.98	0.60	0.42	0.00
Std. Dev.	13.35	3.60	1.07	0.65	0.49	0.00

Table 1: Heuristics Mean and Standard Deviation for Correct Pronoun-Antecedent Pairs in the Validation Set

	Token Distance	Parse Tree Distance	Word Frequency	Antecedent Grammatical Argument	Same Grammatical Argument	Is_Constrained
Mean	15.29	8.71	1.53	0.15	0.41	0.19
Std. Dev.	12.89	4.41	0.93	0.71	0.49	0.50

Table 2: Heuristics Mean and Standard Deviation for Incorrect Pronoun-Antecedent Pairs in the Validation Set

The chosen set of heuristics in decreasing order of significance are*: is_constrained, parse tree distance, antecedent grammatical argument, and word frequency*. The weighting of each heuristic was chosen to maximize performance on the validation set.

Score = 50 * (1/(*parse tree distance*+0.01)) + 10 * (*antecedent grammatical argument*) + *word frequency*

Comparing the scores of antecedent A and antecedent B, the larger one is labeled ′TRUE′ and the other ′FALSE′. However, both are labeled ′FALSE′ if both scores are smaller than 2, which is approximately the lower 2 std. dev. score of the validation set. Also, the pair is regarded as incorrect regardless of the score if *is_constrained* is true. Additionally in page context, the antecedent contained in the Wikipedia URL is labeled ′TRUE′ and the other ′FALSE′. Our results in both contexts are shown on **Table 3**.

	Recall	Precision	F-score
Snippet Context	74.7	67.0	70.6
Page Context	76.7	68.3	72.3

Table 3: Test Results for Snippet Context and Page Context

From the results, we clearly have found a set of useful heuristics for coreference resolution. However, the results could be further improved by using a more statistical approach in optimizing weight parameters. Also, the fact that Page Context performs better than Snippet Context further supports our claim in **3.** and **4.**, which assumes that the topic of discourse is more likely to be the antecedent of the pronoun. Consequently, we could have scraped the Wikipedia articles and attempted to find the topic of discourse for the antecedents that were not found in the URL. Lastly, there inevitably exists limitations in the heuristics scoring approach due to numerous variations in sentence structure and context information. Furthermore, no semantic information is exploited in the heuristic approach. As a result, using deep learning model to capture semantics along with the proposed features in this algorithm would provide a more flexible approach with significantly improved results.