Assignment 2 - Report

Evaluation and Error Analysis

Precision: 0.25142857142857145
Recall: 0.9496402877697842
Values of TP, TN, FP, and FN (respectively): 132, 71, 393, 7

		Ground Truth	
		1 (error)	0 (no error)
Prediction	1 (error)	132	393
	0 (no error)	7	71

Reasons for the False Positives

1. Complex sentences: The grammar has limited capability to recognize some complex sentences, such as those with multiple clauses or subordinate clauses. If a user tries to parse a sentence with these structures that aren't addressed in the grammar, it could be erroneously flagged as having a grammatical error.
   Example:
   - 102 0 It looks nice and has a good message . PRP VBZ JJ CC VBZ DT JJ NN .
     - While the grammar can recognize phrases like NP CC NP, VP CC VP, S CC S, and etc (x CC x). It is likely to miss clauses of different phrase types.
2. Lack of flexibility with commas: The grammar recognizes the comma usages “NP, NP” but the language has many varied examples.
   Examples:
   - 435 0 I keep that in my mind , for ever ! PRP VBP DT IN PRP$ NN , IN RB .
   - 21 0 If not , what do you suggest ? IN RB , WP VBP PRP VB .
   - 58 0 However, we would like to suggest something: RB , PRP MD VB TO VB NN:
3. Inability to recognize multiple verbs: In the grammar, there are no rules that account for two or more verbs that directly follow each other in correct sentences.
   Examples:
   - I have received your letter . PRP VBP VBN PRP$ NN
   - Your time has been stolen . PRP$ NN VBZ VBN VBN .

Reasons for the False Negatives

1. Lack of constraints on the relationship between verb and subject: Without specific constraints on the combination of verbs and subjects, the grammar might approve of sequences that are nonsensical in natural language use. An example of this is a personal pronoun directly followed by a gerund verb.
   Example:
   - 792 1 Be you studing a lot ? VB PRP VBG DT NN .
     - This has a base verb followed by a personal pronoun at the beginning of the sentence.

2. Lack of constraints on determiners: the grammar does not ensure that determiners are used properly, for example, it does not distinguish indefinite vs definite determiners (they are all given the level DT) and does not check for the different rules of both.
   Example:
   - 813 1 I have a big news . PRP VBP DT JJ NNS .
     - This has an indefinite article with a plural noun.

With our current design, is it possible to build a perfect grammar checker?

No, English does not follow perfect CFG (we can not have both precision and recall hold a value of 1), and the original data has mistakes in the POS tagging.

Example of POS tagging mistake in original data:

- 73 0 And always , everybody did it surprisingly well . CC RB , NN VBD PRP RB RB .
  - Should be: CC RB , PRP VBD PRP RB RB . The word “everybody” in this context is a PRP, not an NN.

Assignment 2 - Graduate Extension

Consider how you would approach this task differently if you wanted to also identify colloquial expressions based on the author's location or background. (2-3 paragraphs)

What would you need to do differently?

One way to identify colloquial expressions would be to include customized lexical markers to target the author's location and background. In our task, we did not focus on making customized lexical markers, as we were simply parsing the POS tag sequences given by the original data set. For identification purposes, the lexical markers could be divided into region-specific lexicons. This would help to identify regional dialects and slang used by the author, which can help to hint at their location or background. These lexicons would require substantial information pertaining to the nuances of different dialects. To ensure these nuances are correctly identified, we may need to consult pre-existing dialect-based lexicons (as found within pre-existing research) or perhaps use the assistance of linguists. In terms of our task's current functionality, we may also need to modify the outputted train.tsv with a new column for the predicted region for training purposes, in addition to modify the parsing to use the new lexicon.

What could be kept the same?

The inclusion of a context-free grammar would still be vital to use in identifying colloquial expressions. Without the syntactic rules of sentence construction defined by the context-free grammar, it would be difficult to understand the types of sentences created by the author. By not being able to differentiate between sentence structure, we will also have issues with parsing colloquial language. This is because colloquial language still follows some regular sentence structure, although with some modifications. Using a context-free grammar would also remain helpful with parsing sentences across dialects, as the underlying sentence structure is often quite consistent.

What resources might you need to consult?

Dialectal lexicons already available, for example: https://www.researchgate.net/publication/220746429_Mining_the_Web_for_the_Induction_of_a_Dialectical_Arabic_Lexicon, can be either integrated into the current task's lexicon, or guide in mapping the lexicon. In addition, dialectal dictionaries such as the Dictionary of Caribbean English Usage and the Dictionary of American Regional English are indispensable to both identify colloquialisms and accurately map them and can provide insight into the dialectal variations. Colloquialism and idiom databases are likewise helpful for similar reasons. Finally, linguists trained in dialectical differences in English, or linguists who are experts in specific dialects are an invaluable resource to consult on the task and can additionally offer insight on how to best identify the region, especially when the dialects are closely related.