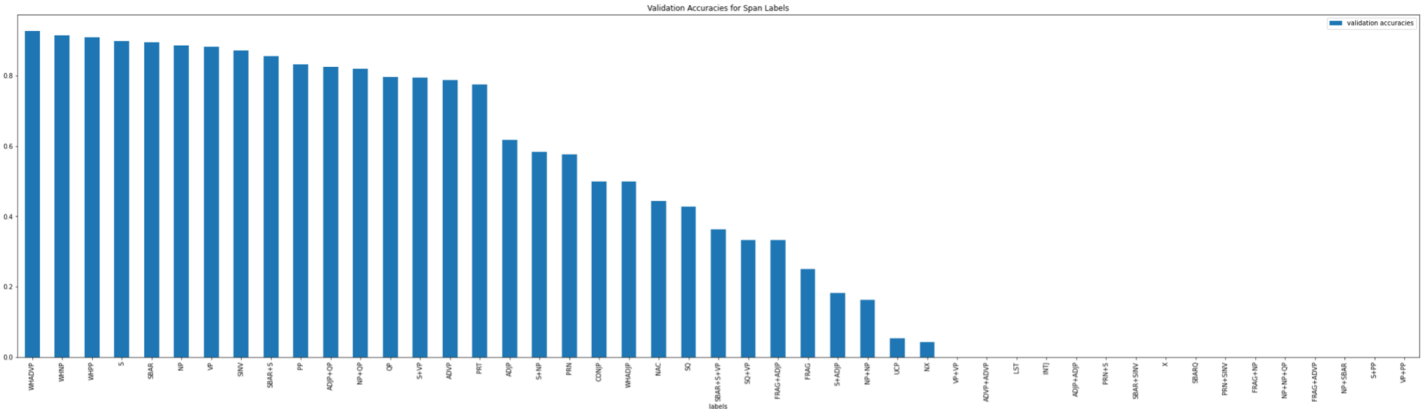
Kristy Lee
CS288 Project 3 Report



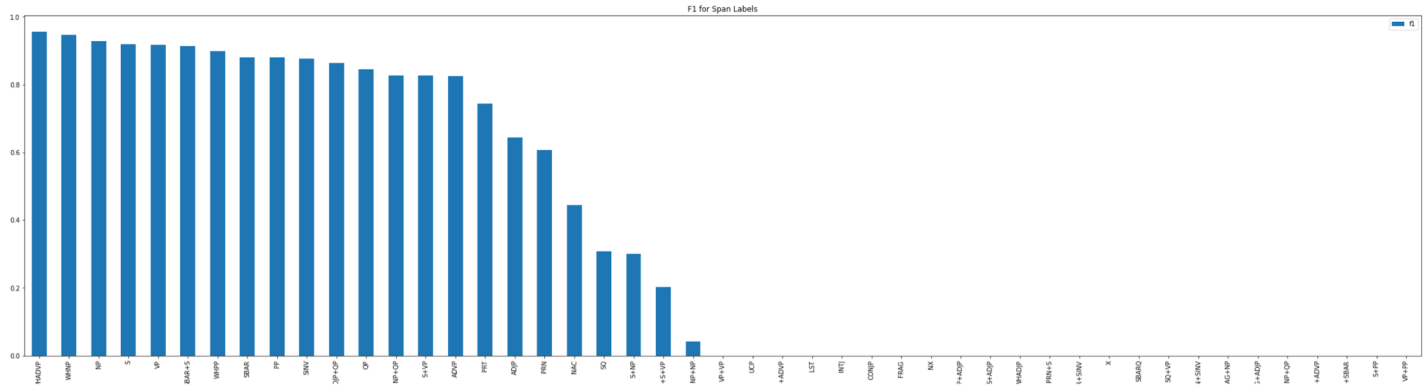
(SPAN LABEL, Accuracy, Count) : [('WHADVP', 0.927710843373494, 83), ('WHNP', 0.9146757679180887, 293), ('WHPP', 0.9090909090909091, 11), ('S', 0.8992832893247831, 2651), ('SBAR', 0.8946700507614214, 788), ('NP', 0.8869654572761047, 12651), ('VP', 0.882678132678132, 4884), ('SINV', 0.8723404255319149, 94), ('SBAR+S', 0.8565573770491803, 244), ('PP', 0.8335051546391753, 3880), ('ADJP+QP', 0.8260869565217391, 23), ('NP+QP', 0.8199233716475096, 261), ('QP', 0.7966101694915254, 177), ('S+VP', 0.7942196531791907, 865), ('ADVP', 0.7874015748031497, 889), ('PRT', 0.7758620689655172, 116), ('ADJP', 0.6174863387978142, 549), ('S+NP', 0.5833333333333334, 12), ('PRN', 0.5769230769230769, 78), ('CONJP', 0.5, 8), ('WHADJP', 0.5, 2), ('NAC', 0.4444444444444444, 27), ('SQ', 0.42857142857142855, 7), ('SBAR+S+VP', 0.36363636363636365, 44), ('SQ+VP', 0.3333333333333333, 3), ('FRAG+ADJP', 0.3333333333333333, 3), ('FRAG', 0.25, 8), ('S+ADJP', 0.18181818181818182, 11), ('NP+NP', 0.16326530612244897, 49), ('UCP', 0.05263157894736842, 19), ('NX', 0.041666666666666664, 48), ('VP+VP', 0.0, 1), ('ADVP+ADVP', 0.0, 2), ('LST', 0.0, 6), ('INTJ', 0.0, 7), ('ADJP+ADJP', 0.0, 3), ('PRN+S', 0.0, 6), ('SBAR+SINV', 0.0, 5), ('X', 0.0, 6), ('SBARQ', 0.0, 2), ('PRN+SINV', 0.0, 1), ('FRAG+NP', 0.0, 1), ('NP+NP+QP', 0.0, 2), ('FRAG+ADVP', 0.0, 1), ('NP+SBAR', 0.0, 1), ('S+PP', 0.0, 1), ('VP+PP', 0.0, 1)]

```
=== Summary ===

-- All --
Number of sentence        =   1700
Number of Error sentence  =     74
Number of Skip  sentence  =      2
Number of Valid sentence  =   1624
Bracketing Recall         =  87.00
Bracketing Precision      =  90.10
Bracketing FMeasure       =  88.53
Complete match            =  29.80
Average crossing          =   0.64
No crossing               =  69.27
2 or less crossing        =  91.56
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   1578
Number of Error sentence  =     59
Number of Skip  sentence  =      2
Number of Valid sentence  =   1517
Bracketing Recall         =  88.41
Bracketing Precision      =  90.93
Bracketing FMeasure       =  89.65
Complete match            =  31.77
Average crossing          =   0.53
No crossing               =  72.58
2 or less crossing        =  93.54
Tagging accuracy          = 100.00
```

The parsing model has achieved 100% tagging accuracy for sentences, and the FMeasure for the parsing model on short sentences of less than or equal to length 40 exceeds 89.0 at 89.65. This indicates that the model does well in correctly classifying constituents with appropriate span labels.

The parsing model has achieved overall good performance on spans with certain word categories and sentence types. I examined the validation dataset and allowed the parsing model to predict labels for specific spans, and some of the span labels that had the highest prediction accuracy in the validation set are WHADVP, WHNP, WHPP, S, SBAR, NP, and VP. WHADVP has accuracy of 0.9277, WHNP has accuracy of 0.9146, WHPP has accuracy of 0.909, S has accuracy of 0.8992, SBAR has accuracy of 0.8946, NP has accuracy of 0.8869, VP has accuracy of 0.88267. While it is pretty expected for S(sentence), NP(noun phrase), and VP(verb phrase) to have high prediction accuracies, I think that it is reasonable and also notable that WHADVP, which refers to “when” that is used as a conjunction, is identified with one of the highest accuracy since “when” is easily identified as being used in conjunction when examining a sentence, such as “He crossed *when* the light turned green.” VP+VP(verb phrase + verb phrase), ADVP+ADVP (adverb phrase + adverb phrase), INTJ (interjection), and multiple other chain labels corresponding to collapsed unary chains, and other span labels have 0% accuracy from the parser, though. I think it makes sense for the chain labels of the collapsed unary chains to have 0% accuracy because it wouldn’t seem common for constituents to correspond to collapsed unary chains so a constituent with a chain label wouldn’t easily be classified with a chain label. Most of the labels that corresponded to 0% accuracy also appeared least frequently in the dev set, which also implies that it wouldn’t be something to commonly classify and thus the accuracy is reasonable.

The span labels follow a similar pattern of ranking from highest F1 score to lowest F1 score (see second page for the plot).



(SPAN LABEL, F1) : [('WHADVP', 0.9565217391304347), ('WHNP', 0.9466192170818504), ('NP', 0.929471032745592), ('S', 0.9195310947744885), ('VP', 0.917319408181027), ('SBAR+S', 0.9130434782608695), ('WHPP', 0.9), ('SBAR', 0.8804861580013503), ('PP', 0.8802846975088968), ('SINV', 0.8770053475935828), ('ADJP+QP', 0.8636363636363636), ('QP', 0.8451612903225807), ('NP+QP', 0.8273092369477911), ('S+VP', 0.8271990018714909), ('ADVP', 0.8246913580246915), ('PRT', 0.744186046511628), ('ADJP', 0.6434195725534307), ('PRN', 0.6071428571428571), ('NAC', 0.4444444444444444), ('SQ', 0.30769230769230765), ('S+NP', 0.3), ('SBAR+S+VP', 0.20253164556962025), ('NP+NP', 0.041666666666666664), ('VP+VP', 0.0), ('UCP', 0.0), ('ADVP+ADVP', 0.0), ('LST', 0.0), ('INTJ', 0.0), ('CONJP', 0.0), ('FRAG', 0.0), ('NX', 0.0), ('ADJP+ADJP', 0.0), ('S+ADJP', 0.0), ('WHADJP', 0.0), ('PRN+S', 0.0), ('SBAR+SINV', 0.0), ('X', 0.0), ('SBARQ', 0.0), ('SQ+VP', 0.0), ('PRN+SINV', 0.0), ('FRAG+NP', 0.0), ('FRAG+ADJP', 0.0), ('NP+NP+QP', 0.0), ('FRAG+ADVP', 0.0), ('NP+SBAR', 0.0), ('S+PP', 0.0), ('VP+PP', 0.0)]