Fitness Metric	Accuracy
COUNT	18.5
RELATIVE	22.0
WEIGHTED	20.4

Table 1: Final accuracy of the metrics

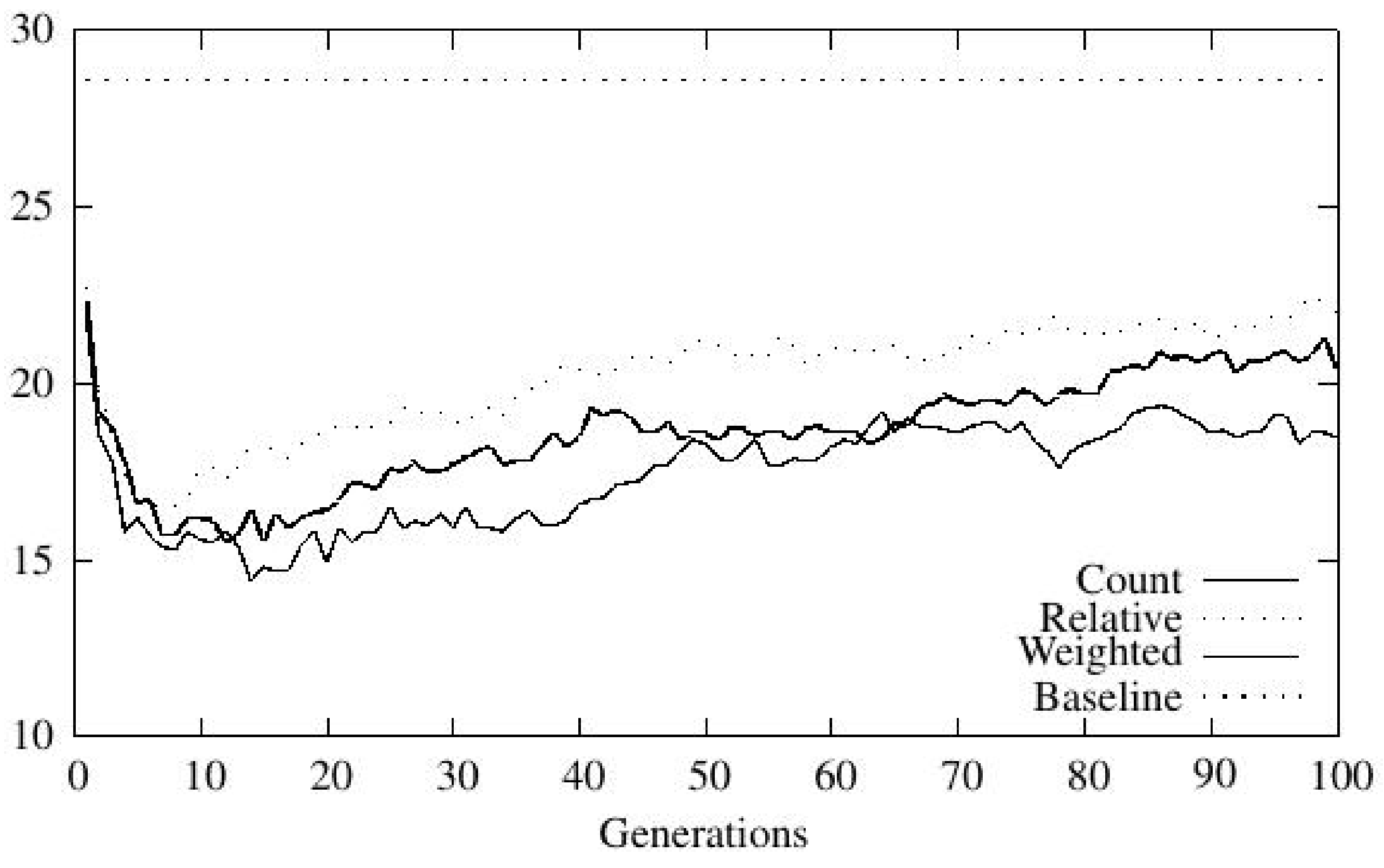

Figure 4: Comparison of fitness metrics

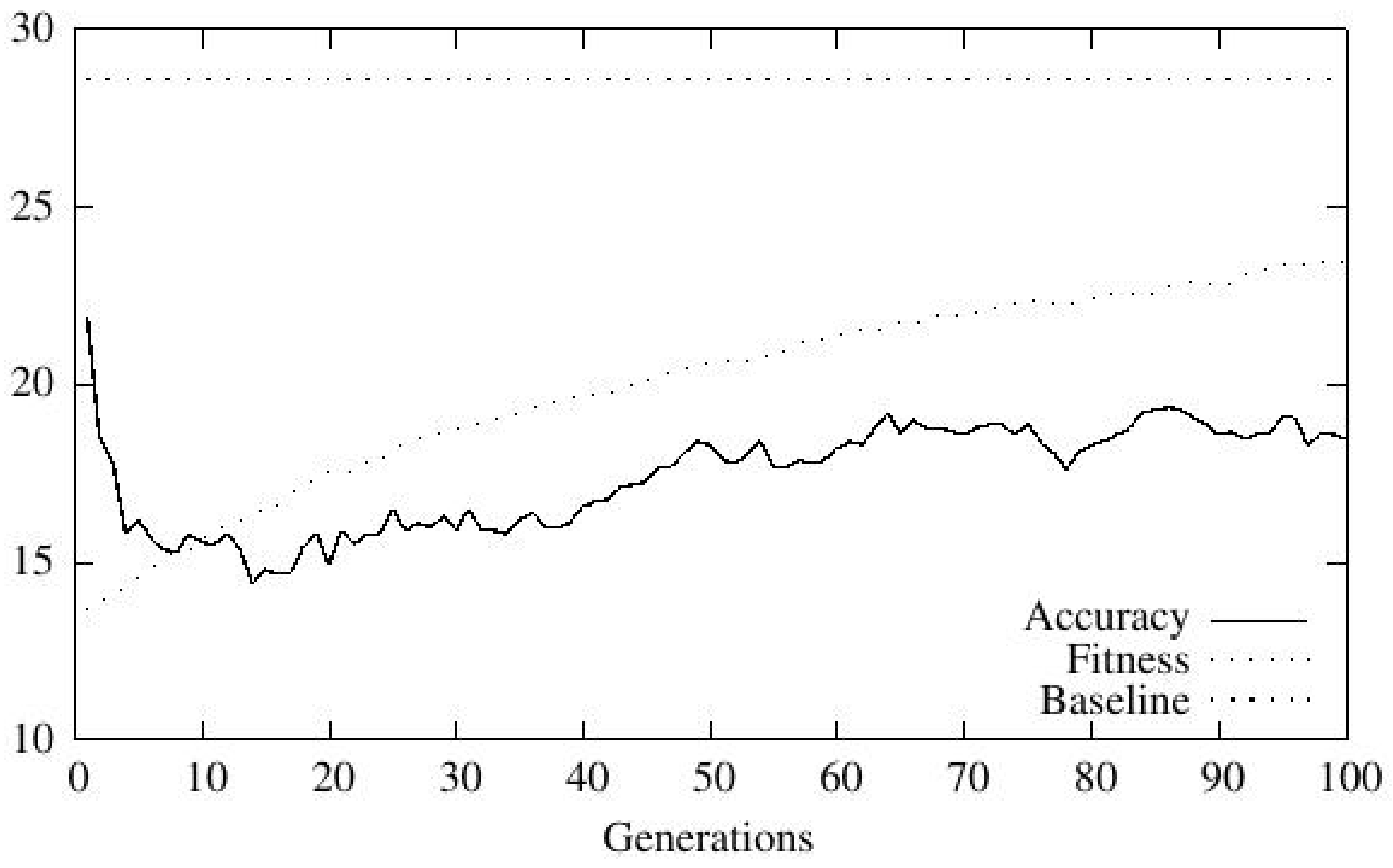

Figure 5: Fitness and accuracy: COUNT

parsing coverage with a GA scheme would correlate with improved category-accuracy.

The end-conditions apply if the parsing coverage for the derived grammar exceeds 90%. Such end-conditions generally were not met; otherwise, experiments ran for 100 generations, with a population of 50 candidates. Because of the heavy reliance of GAs on pseudo-random number generation, individual experiments can show idiosyncratic success or failure. To control for this, the experiments were replicated 100 times each. The results presented here are averages over the runs.

5 Results

5.1 Fitness Metrics

The various fitness metrics were each evaluated, and their final accuracies are reported in Table 1. The results were negative, as category accuracy did not approach the baseline. Examining the average system accuracy over time helps illustrate some of the issues involved. Figure 4 shows the growth of category accuracy for each of the metrics. Pathologically, the random assignments at the start of each experiment have better accuracy than after the application of GA techniques.

Figure 5 compares the accuracy of the category assignments to the GA's internal measure of its fitness, using the Count Spans metric as a point of reference. (The fitness metric is scaled for comparison with the accuracy.) While fitness, in the average case, steadily increases, accuracy does not increase with such steadiness and degrades significantly in the early generations.

The intuitive reason for this is that, initially, the random assignment of categories succeeds by chance in many cases, however the likelihood of accurate or even compatible assignments to words that occur adjacent in the examples is fairly low. The GA promotes these assignments over others, apparently committing the candidates to incorrect assignments early on and not recovering from these commitments. The WEIGHTED and RELATIVE metrics are designed to try to overcome these effects by promoting grammars that parse longer spans, but they do not succeed. Perhaps exponential rather than linear bonus for parsing spans of length greater than two would be effective.

6 Conclusions

This project attempts to induce a grammar from unannotated material, which is an extremely difficult problem for computational linguistics. Without access to training material, logical forms, or other relevant features to aid in the induction, the system attempts to learn from string patterns alone. Using GAs may aid in this process, but, in general, induction from string patterns alone takes much larger data-sets than the one discussed here.

The GA presented here takes a global perspective on the progress of the candidates, in that the individual categories assigned to the individual words are not evaluated directly, but rather as members of candidates that are scored. For a system such as