# Problems with Evaluation of Unsupervised Empirical Grammatical Inference Systems

Menno van Zaanen and Jeroen Geertzen

Dept. of Communication & Information Sciences
Tilburg University
Tilburg, The Netherlands
{mvzaanen, j.geertzen}@uvt.nl

**Abstract.** Empirical grammatical inference systems are practical systems that learn structure from sequences, in contrast to theoretical grammatical inference systems, which prove learnability of certain classes of grammars. All current empirical grammatical inference evaluation methods are problematic, i.e. dependency on language experts, appropriateness and quality of an underlying grammar of the data, and influence of the parameters of the evaluation metrics. Here, we propose a modification of an evaluation method to reduce the ambiguity of results.

## 1 Introduction

Grammatical inference (GI) can be described as the inference or induction of structure from sequences of symbols. We distinguish three sub-fields in the field of grammatical inference: formal GI, empirical GI, and applied GI [1].

Formal GI investigates which classes of grammars can be learned within certain bounds of algorithmic complexity and gives mathematical proofs for this. Empirical GI develops practical systems learning grammars. Often, the underlying (class of the) grammar is unknown. Applied GI is a collection of research that explores or employs GI as a step towards another research goal.

Here, we will review the evaluation methods that are available for measuring the performance of *empirical* GI systems. The sub-field of empirical GI does not allow formal proofs and allows for generic evaluation techniques.

## 2 Current evaluation approaches

Evaluation of GI systems is carried out by applying the system to unstructured data, and evaluating its output. The different methods can be divided into four groups: *looks-good-to-me* approaches analyze the output of GI systems manually. *Rebuilding a-priori known grammars* use, often small, "toy" grammars to generate sequences, which are used as input for the GI system. The output of the system is then compared against the original grammar. The *language membership* method measures the ability to classify sequences based on language

membership. This measures language equivalence (weak equivalence). The performance in this method is expressed by two metrics: *precision*, which shows the effectiveness to decide whether a sequence is in the language or not and *recall*, which measures coverage. Finally, *comparison against a treebank* uses a treebank, a collection of sequences with their derivation, as a "gold standard". The plain sequences (generated by removing the structure from the treebank sequences) are used as input and the output of the GI system is compared against the original structure. [2, 3]

## 3 Problems with current approaches

All evaluation methods have their problems. The *looks-good-to-me* approach is highly subjective. Evaluation performed by the GI system designer is biased and even for external experts it is hard to maintain consistency between systems.

*Rebuilding known grammars* resolves the dependency on experts and biased results, but only small grammars can be tested and scalability is not taken into account. Also, the grammars can be tuned to generate positive results.

The *language membership* methods depend heavily on several design choices. There are different ways to select negative sequences, which has an impact on the results. Similarly, the recall metric requires a sequence generation method, which may also have a large influence.

The *compare against treebank* approach is unbiased with respect to the evaluator and is scalable. However, it still has settings that have a significant impact. This will be discussed in the next section.

The *compare against treebank* and *language membership* methods have most potential. However, the problems of the *language membership* approach require more research, so we will concentrate on the *compare against treebank* approach.

## 4 Evaluating compare against treebank

The *compare against treebank* method uses precision (correctness) and recall (completeness) on tree structures (PARSEVAL [4]) as metrics.

The learned structure is compared against the gold standard, which may contain "trivial" structure. Examples are structures spanning the entire sequence or only a single word. This structure has a impact on the evaluation.

We applied the Alignment-Based Learning system [2] to the ATIS treebank, taken from the Penn Treebank [5] and varied the amount of trivial structure to get an indication of its impact on the evaluation scores. The first column in Table 1 shows the scores on all structure. Columns that are marked with $-e$ discard empty brackets, $-s$ discards brackets spanning the full sentence, and $-w$ discards spans containing a single word only.

Both the micro-average, where counts of correct brackets over the entire treebank are collated, and the macro-average scores are calculated. Micro-averaging is better in showing actual performance by taking bracket distribution per sentence into account.

**Table 1.** Results of Alignment-Based Learning using different evaluation parameters.

| | | -e | -s | -w | -e-s | -e-w | -s-w | -e-s-w |
|---|---|---|---|---|---|---|---|---|
| Macro Recall | 56.18 | 56.18 | 55.73 | 49.19 | 55.73 | 49.19 | 46.79 | 46.79 |
| Macro Precision | 51.13 | 80.57 | 51.07 | 26.26 | 81.75 | 58.24 | 25.40 | 58.57 |
| Macro F-Score | 53.53 | 66.20 | 53.30 | 34.24 | 66.28 | 53.33 | 32.92 | 52.02 |
| Micro Recall | 49.31 | 49.31 | 49.00 | 40.67 | 49.00 | 40.67 | 39.69 | 39.69 |
| Micro Precision | 51.00 | 79.67 | 51.15 | 25.27 | 80.61 | 56.13 | 25.05 | 57.22 |
| Micro F-Score | 50.14 | 60.91 | 50.05 | 31.17 | 60.95 | 47.17 | 30.72 | 46.87 |

The difference between macro- and micro-averaging is substantial and there are major differences with varying amounts of trivial structure. We propose to use the micro-averaged PARSEVAL metrics without trivial structure. This is the most strict evaluation, which, in this case, results in an F-score of 46.87.

## 5 Conclusion

We reviewed empirical GI evaluation approaches, which all have problems. The *compare against treebank* approach is most promising, but it is essential to define the exact settings of the evaluation as they have a major impact in the actual results. We propose to remove all trivial structure and use micro-averaged PARSEVAL metrics. Most published results are difficult to compare and interpret, because the exact evaluation settings are unknown.

## Bibliography


[1] Pieter W. Adriaans and Menno M. van Zaanen. Computational grammatical inference. In Dawn E. Holmes and Lakhmi C. Jain, editors, *Innovations in Machine Learning*, volume 194 of *Studies in Fuzziness and Soft Computing*, chapter 7. Springer-Verlag, Berlin Heidelberg, Germany, 2006.

[2] Menno van Zaanen. *Bootstrapping Structure into Language: Alignment-Based Learning*. PhD thesis, University of Leeds, Leeds, UK, January 2002.

[3] Menno van Zaanen, Andrew Roberts, and Eric Atwell. A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrun Magnusdottir, Anna Samiotou, and Khalid Choukri, editors, *Proceedings of the Workshop: The Amazing Utility of Parallel and Comparable Corpora; Lisbon, Portugal*, pages 58–61, May 2004.

[4] E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of a Workshop—Speech and Natural Language*, pages 306–311, February 19–22 1991.

[5] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.