



Computer Vision and Language
Research Group, University of Leeds

Andrew Roberts & Eric Atwell

*The Use of Corpora for Automatic
Evaluation of Grammar Inference
Systems*

Corpus Linguistics '03 – 29th March



Outline

- Overview of Grammar Inference
- Current evaluation methods:
 - Manual vs Automatic
- Proposed advances
- Implementation issues
- Discussion
- Conclusion



Grammar Inference

- Task of automatically acquiring a type of grammar from a given text.
- Most interesting systems are those which are unsupervised, i.e., learn from raw, unannotated text.
- Examples include:
 - *ABL* (van Zaanen, 2001)
 - *CLL* (Watkinson, 2001)
 - *EMILE* (Adriaans, 1992)
 - *GraSp* (Henrichsen, 2002)



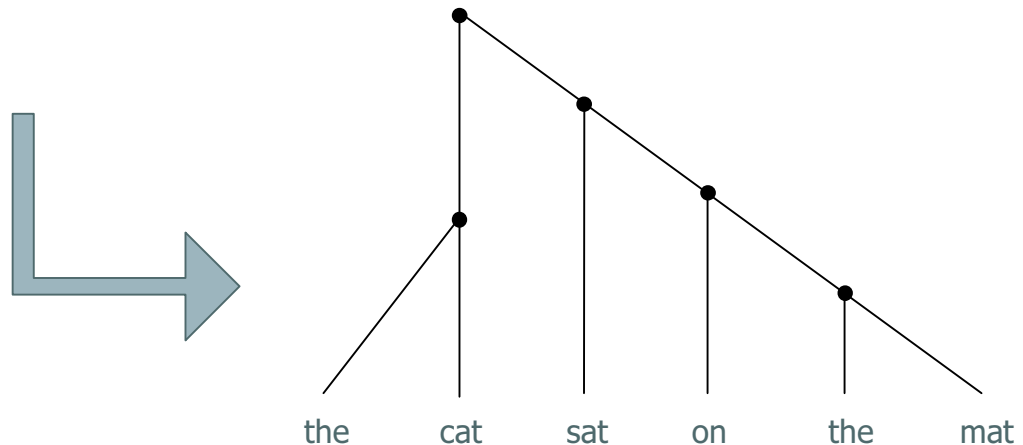
GI Applications

- Psychological/cognitive modelling
- 'Innateness' debate
- Data mining
 - Looking in other datasets, such as bio-informatics
- Finding unique linguistic features that have not yet been discovered by experts

Grammar Inference (Bootstrapping Structure)

- Applies structure to a raw text, e.g., whilst analysing a given text sample, will provide simple annotation using brackets:

- ((the cat) (sat (on (the mat))))





Grammar Inference (Rule Induction)

- Production of grammar rules, e.g, from a given text sample, will induce rules like:
 - $S \rightarrow NP$
 - $NP \rightarrow N VP$
 - etc...
- To use the induced rules, simply plug them into a parser and apply to the text, thus giving structure.



Evaluating GI performance

- Ensuring they are doing the job correctly
- Monitoring improvements of a given system
- Comparing systems



Evaluation

"Looks good to me" approach

- Expert linguist(s) analyse GI output for important linguistic features
- Advantages:
 - Simple
 - Resource efficient
 - Reliable



Evaluation

"Looks good to me" approach

- Disadvantages:
 - Time consuming
 - Relies on subjectivity of expert
 - Possible bias
 - Difficult to compare systems evaluated by experts
 - Can only say which features do and do not exist – no quantitative measures



Automatic Evaluation

- Objectivity and consistency allow for easy comparison between competing GI systems
- Quicker – does not require an expert to perform time consuming analysis



Automatic Evaluation

A 'Gold Standard' corpus

- Compare the results of GI system with an existing treebank:
 1. Make a copy of a treebank and remove all annotation.
 2. Apply GI system to raw treebank text
 3. Compare GI output with the original treebank
- Advantages:
 - Objective
 - Automatic – do not need to be an expert
 - Evaluation results are comparable



Automatic Evaluation

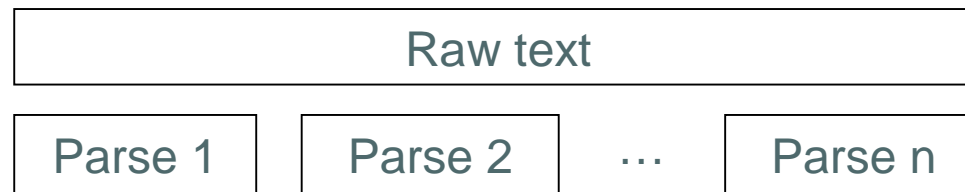
A 'Gold Standard' corpus

- Disadvantages:
 - Treebanks often restricted to particular domain
 - Translation interface between GI output and treebank annotation
 - Is different the same as being wrong?

Automatic Evaluation

A Multi-annotated 'Gold Standard' corpus

- A single treebank that contains more than one set of annotations



- Each parse will be different, but valid
- Compare GI output with each parse
 - Reduces chance of system being discriminated just because it is different from the 'Gold Standard'



Automatic Evaluation

A Multi-annotated 'Gold Standard' corpus

- Problems:

- Different parsers will use their own annotation scheme.

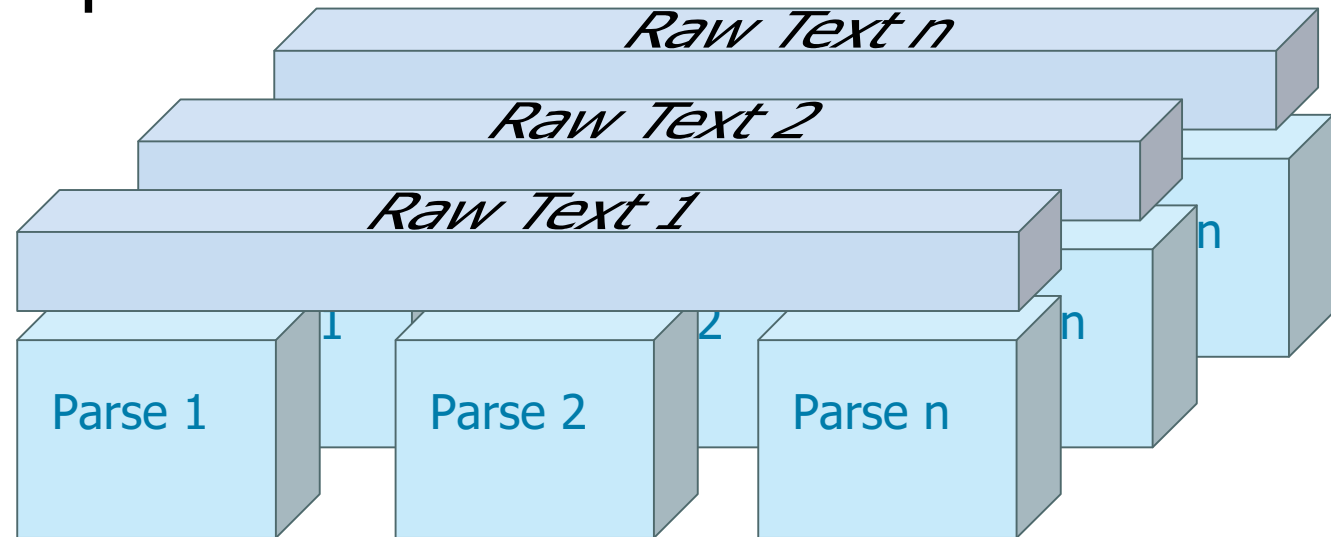
A 'gold standard' annotation scheme would be ideal, but is it feasible?

- What is a match?

Automatic Evaluation

A Multi-annotated-multi-corpora Gold Standard' corpus

- A bit of a mouthful!
- Adds an extra dimension to the last advance by incorporating many corpora





Automatic Evaluation

*A Multi-annotated-multi-corpora
Gold Standard' corpus*

- Advantages:
 - Importantly, this adds multiple genres
 - Makes for a better, more general purpose corpus
- Disadvantages:
 - Adds yet more complexity
 - Requires a lot of time & effort to create



Discussion

- Are automatic and manual approaches equally reliable?
- Researchers are likely to go for the simplest evaluation method
- Create a research project aimed solely at thorough automatic evaluation.
- Build a black box evaluation toolkit, that can be easily added to the end of a GI pipeline



Conclusion

- Despite critical slant, “Looks good to me” is not inferior
- In fact, the goal is to emulate it, but using corpus based methods instead
- Current ‘Gold Standard’ approach is too basic, as it will favour systems that produce results similar to that of the standard itself
- Expand to create a richer corpus



Thank You!

- Thanks for attending this presentation

- **Any Questions?**