A Multilingual Parallel Parsed Corpus as Gold Standard for Grammatical Inference Evaluation

Menno van Zaanen[¶], Andrew Roberts[*] and Eric Atwell[*]

[¶]Tilburg University
Postbus 90153, 5000LE, Tilburg, the Netherlands
mvzaanen@uvt.nl

*University of Leeds
Woodhouse Lane, Leeds, LS2 9JT, U.K.
{andyr, eric}@comp.leeds.ac.uk

Abstract

In this article we investigate how (computational) grammar inference systems are evaluated and how the evaluation procedure can be improved. First, we describe the currently used evaluation methods and look at the advantages and disadvantages of each method. The main problems of the methods are: dependency on language experts, the influence of the annotation scheme of the language data, and the language dependency of the evaluation. We then propose a new method that will allow for an evaluation independently of language and annotation scheme. This method requires (syntactically) structured corpora in multiple languages to test for language independency of the grammatical inference system and corpora structured using different annotation schemes to diminish the influence the annotation has on the evaluation.

1. Introduction

Grammar inference (GI) is focused on the task of inferring or learning grammatical descriptions of a language from a corpus of language examples. Research on grammar inference focuses on showing which (classes of) grammars can be learned and how this can be done. This includes formal learnability research, which identifies, for example, classes of grammars that can be learned within polynomial time and gives mathematical proofs for this. Additionally, linguists (including, among others, formal linguists, psycholinguists, cognitive linguists and computational linguists) concentrate more on natural languages. Discussions and cooperations between the different groups of researchers has led to interesting results (de la Higuera et al., 2003)

On the one hand, formal grammar inference research provides us with solid proof of the learnability of classes of grammars, which might not have any linguistic relevance. On the other hand, researchers from other fields have a harder time actually proving or even showing that a system or approach might actually learn a certain type of language.

In this article we will take a look at the evaluation methods that are available for investigating the performance of grammar inference systems. We will describe the approaches currently in use and discuss their advantages and disadvantages. Based on this, we propose a new evaluation approach. This approach reduces the influence of a specific language or annotation scheme by testing on several different languages and on texts annotated with different schemes.

2. Current evaluation approaches

Several descriptions of grammar inference systems together with some evaluation have been published (see, for example, (Adriaans, 1992; Grünwald, 1994; Nakamura and Ishiwata, 2000; Wolff, 1980)). These and other GI systems have been evaluated using different methods. The evaluation methods used can be divided into three large groups (van Zaanen, 2002).[1] These groups are described below.

2.1 Looks-good-to-me

The GI system is applied to unstructured data. This data can be, for example, linguistic data or it can be generated by a grammar. The output produced by the system is then checked manually for interesting aspects.

This approach as two main advantages. Firstly, only unstructured data is needed. This makes it easy to apply the system on different languages. Secondly, the evaluation can focus on certain specific syntactic constructions. Not only can the output of the GI system be easily searched for a given construction, the input can be tailored to learning it as well.

However, this approach will only provide a useful means of reference if it is done by an independent expert comparing outputs of rival systems. In practice most GI developers have applied looks-good-to-me evaluation to their own systems, rather than perform objectively quantifiable comparisons.

Human evaluation of output is accepted standard practice in Machine Translation evaluation, e.g. (Elliott et al., 2003), where a range of translations may be equally valid. However, this evaluation involves assessments by independent judges, who give an expert assessment of quality of output.

[1] A fourth method, which we call *language membership*, is being used in GI competitions such as Abbadingo, Gowachin, and Omphalos. The learning system must indicate whether a test sentence is a member of the language or not. The correct answers are counted. We will not consider this approach any further, since no explicit grammatical properties are measured.

2.2 Rebuilding known grammars

In this evaluation approach, one or more "toy" grammars are selected beforehand. These grammars are used to generate data, which again is used as input for the GI system. The output (i.e. the grammar or the structured version of the input) is then compared to the original data.

The grammars can be chosen with known properties. These properties can, for example, reflect specific syntactic constructions or be more global, such as context-freeness. Additionally, the evaluation itself can be done automatically, without the need for a language expert.

Because the grammars are chosen or created by hand, this may work for small, artificial languages, but does not scale up to wide-coverage Natural Language grammars. A related problem is that the specific grammars might be tailored to the specific GI systems.[2] On a wider scale, different GI systems aim for different types of grammars or language models, making this an unfair test of systems not geared to generate, for example, small context-free grammars.

The generation part of this approach also poses interesting problems. One has to decide what probability distribution should be assigned to the grammar rules. This decision might influence the learning process. Additionally, all grammar rules should be applied at least once (otherwise the grammar rule cannot be learned) and restrictions may be necessary to limit the sentence length. With respect to emulating natural languages, this comes down to deciding on a language model.

Another problem is that comparing grammars in general is hard. With infinite languages, not all sentences in the language can be compared, which results in a need to compare the generative power of the grammars themselves, which in turn can be quite hard in practice. Note that when the goal is to learn the tree language, this problem is less hard (since the grammar rules themselves can be compared), but not necessarily trivial.

2.3 Compare against treebank

The final approach starts out with an annotated treebank which is selected as a "gold standard". The GI system then infers or rebuilds the structure of the plain sentences extracted from the annotated treebank. The learned, structured sentences are compared against the trees in the original treebank, which measures how well the GI system can find the original structure.

The gold standard is a treebank, that may contain natural language data or tree structures generated by a grammar. This allows for flexibility in the data or grammars used. Different natural languages or data from specific domains can be tested. All GI systems can be adapted to generate structured versions of the input sentences, unlike with the *rebuilding known grammars* approach, where the output of the GI system needs to be a grammar. When a system generates a grammar, the sentences can be parsed, which still results in a structured version of the input sentences. This makes comparing trees a valid option for all systems.

The main problem with this approach is that structured corpora are needed. This may not be a problem when evaluating known grammars, but in the case of natural languages, the underlying grammar is not known. This means that natural language treebanks are needed, which need to be build by hand (or semi-automatically).

3. Problems with current approaches

Although the current approaches provide information on the effectiveness of GI systems and even some standard grammars and test treebanks (Clark, 2001; Klein and Manning, 2002; van Zaanen and Adriaans, 2001) arise, each approach has some problems as described above.

From the existing approaches, the *compare against treebank* approach has most potential. With the *looks-good-to-me* approach, objective evaluation is difficult (especially since often blind evaluation is not performed). The *rebuilding known grammars* approach is too limited because the underlying grammar of natural language data is not currently known. This restricts the application to relatively small artificial grammars.

One of the aims of GI is to achieve generic learning, across a wide range of source language data. Focusing on a specific treebank for comparative evaluations may result in over-training and/or a bias in favor of GI systems developed for a comparable language. Another bold aim of GI is the discovery of new concepts in grammar, or at least valid alternatives to "standard theory". Evaluation by comparison with "received wisdom" will not favor innovation.

Another problem is that, doing evaluation using treebanks is not as simple as one might expect from the discussion above. One needs to decide on several parameters. The metrics that will be used to compute similarity between trees have a huge impact on the final results. Currently, the PARSEVAL metrics[3] are often used (Black et al., 1991), but other measures are of course possible.

Furthermore, we have to keep in mind that to investigate and compare the effectiveness of the wide range of GI systems properly, a robust evaluation method is needed. GI systems are meant to be used on different (natural) languages (and domains), so the evaluation method needs at least to be robust with respect to language. Additionally, since we are considering structure, the annotation of this structure should not be a major factor in the evaluation results. Robustness with respect to annotation should, thus, also be taken into account.

4. Evaluation using a parallel corpus

We propose the use of a parallel-parsed corpus as the new gold standard, as it offers a fairer approach to evaluation, and does not promote over-training as easily (Roberts and Atwell, 2003).

[2] In practice, there are some grammars that are considered "standard" test grammars (Cook et al., 1976; Hopcroft et al., 2001; Nakamura and Matsumoto, 2002; Stolcke, 2003).

[3] The PARSEVAL metrics can compare simple phrase-structure bracket overlap between GI output and Gold Standard phrase-structure parses.

The idea of using a gold standard in itself is not new. There have been similar gold standard approaches to evaluation of parsers (Black et al., 1991), Machine Translation systems (Elliott et al., 2003), and other NLP systems. However, here we try to solve many of the problems of the existing approaches.

4.1 Different languages

Non-English language resources are comparatively rare compared to English ones. We are not only referring to corpora, but to language tools, too. If we are to provide a multi-parsed corpus for each language selected, there must exist a variety of taggers and parsers to achieve this aim.

Fortunately, there are many sizable treebank creation projects under way: Dutch ALPINO treebank (van der Beek et al., 2001), Bulgarian BulTreebank (Osenova and Simov, 2003), UPenn Chinese treebank (Xue et al., 2004), UAM Spanish Treebank (Moreno and López, 1999), NEGRA German treebank (Skut et al., 1997), and many more. These would need to be expanded for our purposes to include parallel parses.

Another aspect to take into consideration is to select a broad range of languages, spanning a variety of language families. This should result in a well balanced corpus. For example, we will obviously have English as one of our candidate languages, which comes from the Germanic branch of the Indo-European family. It would therefore make sense not to include (much data of) another language from this branch such as Dutch or Afrikaans until other language families are represented for better coverage, e.g., Russian from the Slavic branch of the Indo-European family, Arabic from the Semitic branch of the Afro-Asiatic family, Japanese from the Altaic family, etc.

4.2 Different domains

Related to the selection of data from several languages (and language families) is the selection of data from different domains. Current *compare against treebank* evaluations within the field of GI take the ATIS treebank (taken from the Penn Treebank) as gold standard.[4] The problem with this is that the treebank is taken from the limited domain of air travel. A fair evaluation should be done on a treebank taken from a much larger domain or a combination of domains.

4.3 Different annotation schemes

One of the largest and most complex tasks of compiling a parallel corpus (by cherry-picking the most appropriate existing treebanks) will be dealing with the large variety of annotation schemes. There is no standard tagset that is commonly adopted by corpus builders, and so each individual corpus is likely to have its own individual annotation scheme.[5]

For our corpus to be adopted by the GI community for evaluation purposes, these inner variances must be transparent, as few developers would have the patience, or resources, to create their own interfaces for each of the various treebanks within the evaluation corpus. We must ensure, that—at least from the end-users' point of view—there is only a single annotation scheme to deal with.

To achieve this, we must first decide upon the "best" annotation scheme for our entire corpus. For the purposes of grammar induction evaluation, a large and highly specific tagset is not necessary. Next, we must work upon a system for mapping original treebank annotation into the "GI evaluation" annotation. Such an approach has already been successfully applied on a small scale within the AMALGAM project (Atwell et al., 2000).

5. Future work

Clearly, the construction of this corpus is still in its early design stages. It has the potential to be an enormous project in terms of resources required. We can use our current parallel-parsed treebank as a seed for future development. Perfecting the design and required skills for compiling a single language, large-scale, multi-treebank is an ongoing process, which entails selecting suitable candidate treebanks, parsers and an annotation scheme. Once this multi-treebank is complete, the next stage will be to apply the same principles for additional languages.

With respect to the practical evaluation using a multi-lingual, parallel corpus, one would like to allow easy access to this data. Preferably, an (operating system independent) software suite should be developed that applies the GI system to the plain sentences of the treebank and compares the output against the structures found in the treebank.

It may prove difficult to automatically compare GI output against Gold Standard trees in all cases, so a fall-back may be to use human "looks-good-to-me" assessment; but in this case the judges are constrained to assess how close the GI output is to the example parse, as in Machine Translation evaluation experiments (Elliott et al., 2003).

The suite should be flexible with respect to different languages, domain specific sub-corpora, annotation schemes and evaluation metrics. This flexibility is needed, for example, when a GI system is computationally intensive and can only be applied to a limited amount of data.

6. Conclusion

In this article, we have investigated the current evaluation approaches that are applied to grammatical inference systems. The approaches can be classified in three groups: *looks-good-to-me*, *rebuilding known grammars*, and *compare against treebank*. Each of these approaches have some advantages, but also disadvantages.

[4] Recently, people have started to use the WSJ treebank for evaluation, but this does not entirely solve the problem (Klein and Manning, 2002; van Zaanen, 2002).

[5] Different languages may, for example, have the need for different part-of-speech tags. Design issues like this influence the annotation of the corpus. Additionally, a treebank may be structured with respect to different syntactic phenomena.

We propose to use a multi-lingual, parallel-parsed corpus as the basis of the evaluation. By applying the system to multiple languages within different domains, the language and domain independency of the GI system is evaluated, while the evaluation against the different parses of the sentences diminishes the impact of the used annotation scheme. In other words, it extends the *compare against treebank* approach in that it also measures the amount of language and annotation scheme independency of the GI system.

References

Adriaans, Pieter Willem, 1992. *Language Learning from a Categorial Perspective*. Ph.D. thesis, University of Amsterdam, Amsterdam, the Netherlands.

Archer, D., P. Rayson, A. Wilson, and T. McEnery (eds.), 2003. *Proceedings of the Corpus Linguistics 2003 conference; Lancaster, UK*.

Atwell, E., G. Demetriou, J. Hughes, A. Schiffrin, C. Souter, and S. Wilcock, 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal, International Computer Archive of Modern and medieval English*, 24:7-23.

Black, E., 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, 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In Proceedings of a Workshop--Speech and Natural Language.

Clark, Alexander, 2001. Unsupervised induction of stochastic context-free grammars using distributional clustering. In *Proceedings of the ACL Workshop on Computational Natural Language Learning; Toulouse, France.*

Cook, Craig M., Azriel Rosenfeld, and Alan R. Aronson, 1976. Grammatical inference by hill climbing. *Informational Sciences*, 10:59-80.

de la Higuera, Colin, Pieter Adriaans, Menno van Zaanen, and Jose Oncina (eds.), 2003. *Proceedings of the ECML Workshop and Tutorial on Learning Context-Free Grammars; Dubrovnik, Croatia*.

Déjean, Hervé, 2000. ALLiS: a symbolic learning system for natural language learning. In Claire Cardie, Walter Daelemans, Claire Nédellec, and Erik Tjong Kim Sang (eds.), *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop; Lisbon, Portugal*. Held in cooperation with ICGI-2000.

Elliott, Debbie, Anthony Hartley, and Eric Atwell, 2003. Rationale for a multilingual aligned corpus for machine translation evaluation. In (Archer et al, 2003), pages 191-200.

Grünwald, Peter, 1994. A minimum description length approach to grammar inference. In G. Scheler, S. Wernter, and E. Riloff (eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language*. Berlin Heidelberg, Germany: Springer-Verlag, pages 203-216.

Hopcroft, J.E., R. Motwani, and J.D. Ullman, 2001. *Introduction to automata theory, languages, and computation*. Reading:MA, USA: Addison-Wesley Publishing Company.

Klein, Dan and Christopher D. Manning, 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL); Pennsylvania:*PA, USA.

Moreno, A. and S. López, 1999. Developing a Spanish tree bank. In *Proceedings of the ATALA Treebank Workshop (Journés ATALA, Corpus annotés pour la syntaxe); Paris, France*.

Nakamura, K. and M. Matsumoto, 2002. Incremental learning of context-free grammars. In Pieter Adriaans, Henning Fernau, and Menno van Zaanen (eds.), *Grammatical Inference: Algorithms and Applications (ICGI); Amsterdam, the Netherlands*, volume 2482 of Lecture Notes in AI. Berlin Heidelberg, Germany: Springer-Verlag.

Nakamura, Katsuhiko and Takashi Ishiwata, 2000. Synthesizing context free grammars from sample strings based on inductive CYK algorithm. In Arlindo L. Oliveira (ed.), *Grammatical Inference: Algorithms and Applications (ICGI); Lisbon, Portugal*, volume 1891 of Lecture Notes in AI. Berlin Heidelberg, Germany: Springer-Verlag.

Osenova, Petya and Kiril Simov, 2003. The Bulgarian HPSG Treebank: Specialization of the annotation scheme. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories; Växjö, Sweden*.

Roberts, Andrew and Eric Atwell, 2003. The use of corpora for automatic evaluation of grammar inference systems. In (Archer et al., 2003), pages 657-661.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans. Uszkoreit, 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97); Washington:DC, USA.*

Stolcke, A., 2003. Boogie. ftp://ftp.icsi.berkeley.edu/pub/ai/stolcke/software/boogie.shar.Z.

Stolcke, Andreas and Stephen Omohundro, 1994. Inducing probabilistic grammars by bayesian model merging. In *Proceedings of the Second International Conference on Grammar Inference and Applications; Alicante, Spain.*

van der Beek, Leonoor, Gosse Bouma, Robert Malouf, and Gertjan van Noord, 2001. The Alpino dependency treebank. In Mariët Theune, Anton Nijholt, and Hendri Hondorp (eds.), *Computational Linguistics in the Netherlands 2001--Selected Papers from the Twelfth CLIN Meeting; Enschede, the Netherlands, Language and Computers: Studies in Practical Linguistics*. Amsterdam, the Netherlands: Rodopi.

van Zaanen, Menno, 2002. *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, University of Leeds, Leeds, UK.

van Zaanen, Menno and Pieter Adriaans, 2001. Alignment-Based Learning versus EMILE: A comparison. In *Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC); Amsterdam, the Netherlands.*

Wolff, J. Gerard, 1980. Language acquisition and the discovery of phrase structure. *Language and Speech*, 23(3):255-269.

Xue, Nianwen, Fei Xia, Fu-Dong Chiou, and Martha Palmer, 2004. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 10(4):1-30.