A Study on the Portability of a Grammatical Inference System

Hsue-Hueh Shih[1] and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, England

1 Abstract

This paper presents a study on the portability of our grammatical inference system called CAGC (Computer Assisted Grammar Construction). The CAGC system has been developed [1] to generate broad-coverage grammars for large natural language corpora. It utilises both an extended Inside-Outside algorithm [2] and an automatic phrase bracketing (AUTO) technique [3], which is designed to provide the extended algorithm with constituent information during learning. The system is firstly trained and tested on the Wall Street Journal (WSJ) corpus, and then ,for the study of its portability, it is moved onto the Brown Corpus to infer a Brown grammar. The experimental results shown in this paper demonstrate that the CAGC inference technique as well as the initial grammar used in the system are transferable to the new corpus.

2 Introduction to Grammatical Inference

Grammar is a crucial component in most natural language processing systems because it bounds the range of constructions which can be handled. However, the conventional method of manual grammar construction is labour intensive, time consuming and often leads to errors caused by unwanted rule interactions. In addition, manually-developed grammars often rely on the assumption that all input sentences are well-formed. Consequently, these grammars have limited coverage on naturally occurring corpora. To go beyond this traditional approach, more practical and robust techniques for grammar construction become necessary.

With the increasing availability of large naturally occurring text corpora in machine readable form, it has become possible to infer linguistic knowledge directly from regularities that appear in sentence samples. The application of techniques for inferring syntactic information in such a way is termed Grammatical Inference (GI). Among recently developed GI techniques, the Inside-Outside algorithm [4] shows potential for the inference of stochastic context-free grammars. However, its practical use in Natural Language Processing is limited by both its high computational complexity and there being no guarantee of convergence to a local optimum which is linguistically motivated.

Recent improvements to this technique have included supervised training [2] to accelerate the inference process and the use of an Explicit-Implicit technique [5] employing a hybrid initial grammar to bias the inference process towards linguistically meaningful solutions. Nevertheless, supervised training re-introduces the problem of labour intensiveness, since the required treebank must be manually annotated. Alternatives must be sought to alleviate this manual load as well as provide useful constituent information for training. The CAGC system, whose portability is examined and will be shown later in this paper, integrates a heuristic-based surface bracketing with the Explicit-Implicit technique to complement the inference process.

[1] Hsue-Hueh Shih now works in the Department of Foreign Languages and Literature, National Sun Yat-sen University, Taiwan. E-mail: hsuehueh@mail.nsysu.edu.tw

The Overview of the CAGC system is given in the next section, which is followed by a system evaluation on the WSJ in Section 4 and the portability study using Brown Corpus in Section 5. Conclusions are drawn in Section 6.

3 Overview of the CAGC system

The CAGC system takes advantages of both heuristic and stochastic approaches. Heuristic knowledge provides powerful and important constraints to the system, whereas stochastic information deals with situations which are too complex or too trivial for heuristic rules to handle. A block diagram of the system is shown in Figure 1.

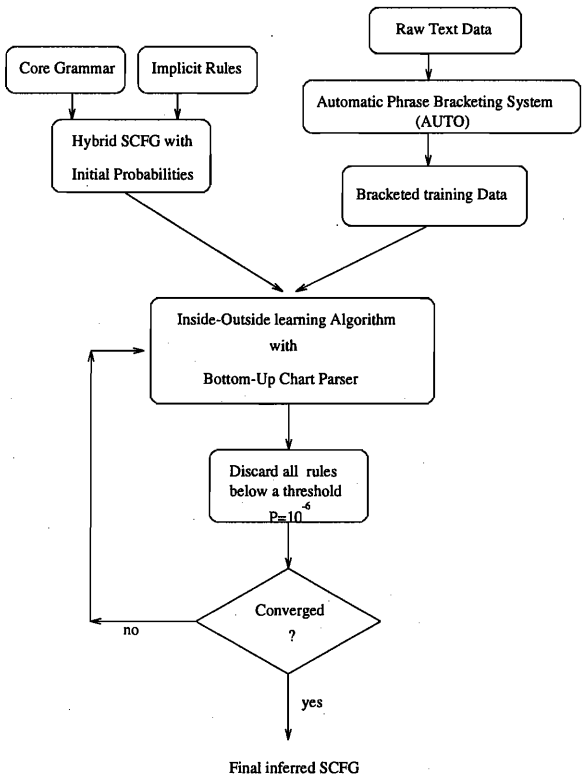

Figure 1: A Block Diagram of the CAGC System

The first part of the system falls into two stages: construction of an initial SCFG and phrase-bracketing of the raw text data. In the second part of the system, a grammar is inferred by utilising the Inside-Outside algorithm to re-estimate the initial SCFG from the bracketed text data. The initial SCFG is derived from the hand-written core grammar (explicit part), which forms a skeleton of the SCFG, and a set of CF rules (implicit part) which consists of all possible rules that do not appear in the core grammar but are nevertheless linguistically plausible. The explicit and implicit rules are then integrated into a hybrid SCFG along with an appropriate set of initial probabilities. Details of the grammar development and the calculation of the initial probabilities are described in [3]. The AUTO bracketing technique utilises heuristic knowledge to bracket the raw text data in a way which integrates top-down and bottom-up approaches. The training set augmented by this derived constituent information provides the additional constraints to the grammar re-estimation process in the second part of the CAGC system.

In the second part of the CAGC system, the Inside-Outside learning procedure, incorporating a bottom-up chart parser [6], iteratively re-estimates the probabilities of the production rules. The updated probabilities are calculated according to the weighted frequency counts of the rules used in parses licenced by the grammar and generated at the previous iteration. At the end of each iteration, the rules with probabilities falling below a pre-defined threshold are discarded. The re-estimation process continues until either the change in the total log probability between iterations is less than a minimum or the number of iterations reaches a maximum. The final inferred grammar is generated when either criteria is met.

4 System Evaluation on the WSJ Corpus

1500 training and 500 test data were chosen from the Wall Street Journal(WSJ) text corpus. There is no explicit limitation on their length, and the average length of data sentences is around 13 words. Instead of lexical entries, parts-of-speech (POSs) are used in our experiments to reduce computation. Original 48 WSJ POSs were manually subcategorized into 59 in order to capture more detailed syntactic information. Detailed subcategorization is stated in [3].

Table 1 shows the performance of the inferred WSJ grammar, when compared with an inferred grammar supervised by Penn treebank. It records the number of SCF rules which survived after training, the number of test sentences which can be parsed by the grammar, and the performance on three metrics that are often used to evaluate NLP systems [7]. Recall is the percentage of standard bracketings (in Penn treebank) present in our experimental output of the same sentence. This metric indicates the closeness between the evaluated grammar and the Penn treebank. Precision is the percentage of the bracketings in our output present in the Penn treebank sentence. A crossing error is defined as the partial overlap between a bracket pair (one generated from our experiment and the other from the treebank) and Crossings are the average number of crossing errors in a sentence.

Grammar Types	**PENN_Trained**	**AUTO_Trained**
Rules Remaining After Training	21.29% (6029/14736)	18.54% (2733/14736)
Sent. Parsed	97.80% (489)	97.20% (486)
Recall	84.65%	84.66%
Precision	64.06%	62.50%
Crossings	1.92	2.14

Table 1: Performance of WSJ Grammars Trained on PENN treebank or AUTO-bracketed data

Figures in Table 1 demonstrate that the CAGC inference technique is able to generate a high coverage grammar with good accuracy in phrase bracketing, and AUTO is capable of providing useful and competitive bracketing information during training phase.

As the data used in the experiment were manually tagged, it is desirable to integrate an automatic tagger into the CAGC system, so that the system no longer requires any pre-tagged data for its training. For this reason, the Acquilex tagger [8] was trained on a subset of the WSJ corpus, and then integrated into the CAGC system as a front-

end. Table 2 shows the performance of the CAGC system using Acquilex-tagged data. Note that this experiment was carried out on increased training (4000) and test (1500) sets, which results in a better performance, when compared with the corresponding figures in Table 1 before the tagger is employed. From Table 2, one can see that the performance of the inferred grammar degrades as the tagger is introduced into the system. This degradation is due to the 7% error rate of the tagger.

System	Manually-tagged	Acquilex-tagged
Recall	86.56%	83.06%
Precision	64.25%	61.79%
Crossings	1.93	2.31

Table 2: The CAGC System Performance Using a Tagger as the Front-end

5 Portability Evaluation on the Brown Corpus

The portability of the CAGC system is investigated using the Brown corpus. Similar sizes of 4000 training and 1500 test data were collected for this experiment. These data are given consistent POSs by the Acquilex tagger. The hybrid initial grammar is directly transfered from the WSJ task. The CAGC system re-estimates the parameters of the grammar iteratively, according to the Brown training data which is AUTO bracketed in advance. The final inferred Brown grammar is generated and then used to analyse the test data. Table 3 shows the performance of the inferred Brown grammar when compared with that in the WSJ task.

Inferred Grammar	WSJ	Brown
Recall	83.06%	79.04%
Precision	61.79%	57.64%
Crossings	2.31	3.10

Table 3: Performance of the Inferred Brown Grammar on 1500 Test data

As can be seen, the overall performance on the three metrics degrades in the Brown task. Recall and Precision are both down 4%, whereas Crossings increase to 3 errors for a sentence. In order to account for this degradation, two additional experiments on the accuracy of Acquilex and AUTO are carried out (the details of these experiments can be seen in [9]). The first experiment on the tagging performance of the Acquilex tagger shows that the tagging accuracy decreases from 93% in the WSJ to 91% in the Brown tasks. This is because the tagger was trained on WSJ, and therefore the proportion of the unknown words to the tagger was larger in the Brown data. This situation can be easily improved by using a larger set of data from different copora as training material for the tagger.

The second experiment on the bracketing accuracy of AUTO shows there is a 6% decrease in both Recall and Precision metrics and Crossing errors increases 0.6 for a sentence. As AUTO works on the POS sequences, it is believed that this is caused partly by the decreasing accuracy of the tagger and partly by the fact that it is designed originally for the WSJ task. AUTO will need to be re-tuned to meet the requirement of task-independency.

From the experiments shown above, it is felt that the 4% decrease in the overall performance of the inferred grammar is mainly caused by the decreasing accuracy in

both Acquilex and AUTO. Therefore, making them more task-independent becomes a key issue on improving the portability of the CAGC system. Nevertheless, the hybrid initial grammar is believed to be transferable to the new corpus, since its core part is designed to capture important general syntactic structures in English grammar and its implicit part will be shaped to target the corpus-dependent structures.

6 Conclusions

Portability is a significant issue and usually involves a large amount of manual work in most grammar-based systems. A grammar designed for one corpus may not properly apply to another corpus and ,therefore, modifying the grammar manually is often required when moving from one application to another. In this paper, the CAGC system shows its potential in alleviating this problem. From the experimental results shown, it is believed that inference technique and the initial hybrid grammar are transferable to the new corpus, and the portability of the system can be improved if two of the CAGC compoents, the Acquilex tagger and the AUTO phrase bracketing technique, are made more task-independent.

References

[1] H-H. Shih, S.J. Young, and N.P. Waegner. An inference approach to grammar construction. *Computer Speech and Language*, 9:235–256, 1995.

[2] F. Pereira and Y. Schabes. Inside-Outside re-estimation for partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, June 1992.

[3] H-H. Shih and S.J. Young. A system for computer assisted grammar construction. Technical Report TR.170, Engineering Department, Cambridge University, England, June 1994.

[4] J.K. Baker. Trainable grammar for speech recognition. In *Speech Communication Papers for the 97th Meeting of the acoustical Society of America (D. Klatt and J. Wolf, eds)*, pages 547–550, 1979.

[5] E. Briscoe and N. Waegner. Robust stochastic parsing using the inside-outside algorithm. In *AAAI Symposium on Statistic Applications to Natural Language*, June 1992.

[6] G. Gazdar and C. Mellish. *Natural Language Processing in PROLOG.* Addison-Wesley, 1989.

[7] H.S. Thompson. Parseval workshop. In *ELSNews Vol.1(2)*, 1992.

[8] D. Elworthly. Part-of-speech tagging and phrasal tagging. Technical Report Acquilex-II Working Paper 10, Computer Laboratory, Cambridge University, England, 1993.

[9] H-H. Shih. *Computer Assisted Grammar Construction.* PhD Thesis. Engineering Department, Cambridge University, England, 1995.