Experimental acquisition of grammar from early reader books

Stephen Muggleton,
Oxford University Computing Laboratory,
Parks Road,
Oxford, OX1 3QD,
United Kingdom.

Abstract

Inductive Logic Programming has demonstrated strengths over other machine learning techniques on natural language data. Domains include discovery of rules governing the formation of the past tense of verbs in English. In this paper we describe initial results of a pilot study aimed at acquiring English natural language grammars directly from the text of a children's book. The training and test sentences are taken from Books 1 and 2 of the Ladybird "Read with me" series. Cross-validation results for Book 1 are compared with hold-out results based on testing the grammar learned from Book 1 on Book 2 sentences. In the experiment, the legal sentence predicate is learned given complete background knowledge of relevant phrase structures. Both cross-validation and hold-out results for the experiment are consistent with existing sample complexity predictions for learning from positive examples only. However, hold-out testing on Book 2 violates the usual assumption in Probably Approximately Correct learning that training and test examples are drawn from the same distribution. Whereas this assumption is reasonable in the case of scientific theory formation, the results underline the fact that it is unreasonable in the case of graded tutorial examples. A planned extension of this experiment is aimed at the challenging task of incremental learning from books 2 to 16.

1 Introduction

It is increasingly clear that natural language (NL) research is in need of machine learning (ML) techniques. On the one hand the irregularities of natural language have overwhelmed attempts to manually encode complete and correct grammars. On the other hand vast tagged corpora of natural sentences are now available on

CD ROM. Some successes have been demonstrated in applying ML techniques to tag prediction within corpora [2]. However, various problems lie in the path of using learning in the way humans do for acquiring natural language grammars. From the NL side, sources of natural language tend to contain overly complex sentences. Consider the following typical sentence from the editorial leader of the New Scientist magazine (*No. 2029, 11th May 1996*).

> It is no surprise that the rogue regimes of countries such as Iraq can recruit scientists with similar talents into biological warfare programs.

From the point of view of a machine (or human) learner with limited background knowledge such a sentence would be unreasonably complex as initial training material. It should also be noted that almost all standard applied ML techniques [4] are attribute-based, and thus poor at dealing with sequence information. Although inductive grammar acquisition has been a strong topic of theoretical research since early papers by Gold and others [3, 1], this has not led to a stream of applied results.

Inductive Logic Programming (ILP) has been demonstrated to have advantages over other ML techniques on certain constrained NL problems [5]. However, to the author's knowledge, ILP has not previously been used for learning NL grammars directly from existing text. In this paper an initial attempt is made at this problem. However, rather than using adult-level text, we take examples from the first two books in the children's Ladybird "Read with me" book series [8, 9]. This has the advantage of providing a source of natural data, with sufficiently low complexity for ILP to be effective. The analogies with human grammar acquisition are clear. High cross-validation results for Book 1 have led the author to plan a series of experiments aimed at incrementally learning from Books 2 to 16 of the Ladybird "Read with me" series. However, the initial results in this paper have already demonstrated a mismatch in this task with standard theoretical assumptions. Whereas the PAC model assumes that training and test examples are drawn from the same distribution, this assumption is violated in the case of graded tutorial examples, such as those provided by the Ladybird books. The clash in assumptions is brought out experimentally in this paper by the mismatch between cross-validation results on Book 1 and hold-out results of testing the Book 1 derived grammar on examples in Book 2.

In the Book 1 experiment it was necessary to introduce a substantial amount of background knowledge concerning phrase structure. Subsequent experiments on Books 2 to 16 are expected to require knowledge revision techniques to augment this initial background knowledge. The Book 1 experiment also acts as a further bench-test of the positive-example-only learning technique used in Progol4.2 [6]. In [6], the assumptions of the theoretical model developed are tested on an artificial grammar with randomly chosen data. The present paper can be taken as a test of whether these assumptions can be interpreted in a reasonable way for real world data.

This paper is structured as follows. Section 2 introduces the experiment aimed at learning a grammar for the Ladybird "Read with me" Book 1. In Sections 2.1, 2.2, 2.3 and 2.4 the experimental hypothesis, materials and method are described. Lastly in Section 3 we discuss the results of the experiment together with various issues for planned research.

2 The experiment

In a previous paper [6] a Bayesian method for learning from positive data was described and analysed. The Bayesian model assumes distributions $D_{\mathcal{H}}$ and D_X are supplied over the hypothesis and instance spaces respectively. A randomly chosen target T drawn from $D_{\mathcal{H}}$ is assumed to have a complexity of $-log_2 D_{\mathcal{H}}(T)$ bits. The error of an hypothesis H is $D_X(H \setminus T) + D_X(T \setminus H)$. The approach described in [6] considers the case of the maximum posterior probability hypothesis given positive examples only. For this hypothesis, on average the error-rate given m randomly chosen positive instances is shown to be at most $\frac{\mathrm{SZ}(T)+2}{m}$.

The problem with applying this result directly to cross-validation results for the Ladybird "Read with me" Book 1 sentences are as follows.

$D_{\mathcal{H}}$. It is necessary to assume a prior distribution over target hypotheses. For simplicity it is assumed that $D_{\mathcal{H}}$ is consistent with an encoding of hypotheses which gives $sz(H)$ to be simply proportional to the number of literals in H with a constant of proportionality of at least 1.

D_X. While the positive examples are simply the sentences in Book 1, it is necessary to assume a distribution D_X to evaluate error-rate over unseen negative examples. The simplest assumption is that D_X decays exponentially with increasing length of sentences. However, it will easily be seen that there is a very low probability that a randomly constructed word sequence is grammatically well-formed. In tests, this probability turned out to be around 3 in 10000. This gives near perfection to the majority class predictor which states that all sentences are ungrammatical. To avoid this anomaly stratified sampling was used to produce a distribution which draws positive and negative examples with equal probability. Whereas the training examples are the positive example sentences from Book 1, testing involves additional negative examples generated randomly by a Stochastic Logic Program [7].

2.1 Experimental hypothesis

Ladybird "Read with me" Book 1 contains a training set of 36 sentences, including 3 repeats. The experimental hypothesis is that given all 36 examples the error of Progol 4.2's proposed theory H will be be at most $\frac{\mathrm{SZ}(H)+2}{36}$.

	A	$\overline{A}$
P	35	0
$\overline{P}$	1	36

Overall accuracy $= 98.61\% \pm 1.38\%$.
χ^2 probability < 0.0001.

Figure 1: Cross validation results for Book 1

2.2 Materials

The example sentences from Book 1 are given in Appendix A. For the purposes of cross-validation of Book 1, negative examples were drawn randomly using a Stochastic Logic Program [7] conditioned on the positive examples from Book 1 (see discussion of D_X in previous section). Appendix B shows the 36 random negative examples. The background knowledge used is given in Appendix C.

For the hold-out results, positive test sentences consisted of the 38 sentences of Book 2. Again 38 negative examples were constructed randomly using a Stochastic logic program.

2.3 Method

Training and testing were carried out using Progol4.2 [6]. Testing involved leave-one-out on the 36 Book 1 sentences and 36 randomly generated negatives. For each of the 72 trials Progol4.2 received only the positive segment of the data for training.

Hold-out testing was carried out on the 38 sentences of Book 2 and 38 randomly generated negatives.

2.4 Results

Each of the 72 cross-validation training runs took under 1 minute of CPU time on a SPARC 10. The cross validation results for Book 1 are given in Figure 1. These include the Actual versus Predicted contingency table, overall accuracy and χ^2 probability. The single omission error indicated in Figure 1 was the sentence "No no". Figure 2 shows the definite clause grammar learned from all 36 examples in Book 1. In Figure 2 the predicates have the following interpretations.

s/2. Legal sentence.

conj/2. Conjunction.

np/2. Noun phrase.

```
s([no,sam,no],[]).
s([no,no],[]).
s(A,B) :- s(A,C), conj(C,D), s(D,B).
s(A,B) :- np(A,C), vp(C,D), np(D,B).
```

Figure 2: Grammar learned from sentences in Book 1

	A	$\overline{A}$
P	31	0
$\overline{P}$	7	38

Overall accuracy $= 90.79\% \pm 3.32\%$.
χ^2 probability < 0.0001.

Figure 3: Hold-out results for training on Book 1, testing on Book 2

vp/2. Verb phrase.

Hold-out results based on training with Book 1 and testing with Book 2 (augmented by randomly generated negatives) are shown in Figure 3. Again these consist of the contingency table, overall accuracy and χ^2 probability. The seven errors of omission from Book 2 are as follows.

1. Kate likes the trees and the dragon.
2. Tom likes the trees and the dragon.
3. Here is Sam the dog.
4. Sam the dog is here.
5. Tom and Kate like the trees and the dragon.
6. I like the toy trees.
7. I like Kate and Tom and Sam.

Book 2 makes use of two new nouns not used in Book 1. These are "dragon" and "trees". However, when the background knowledge for Book 1 is augmented with these new nouns, only omission 6 above is corrected. The remaining six errors of omission are due to inadequate phrase structure knowledge in the background. In particular, this occurs with the new use of compound noun phrases such as

"the trees and the dragon" as well as "Kate and Tom and Sam" and "Sam the dog".

The experimental hypothesis was that the error of Progol 4.2's proposed theory H will be be at most $\frac{\mathrm{SZ}(H)+2}{36}$. It was assumed that $\mathrm{sz}(H)$ was proportional to the number of literals in H, where the constant of proportionality is at least 1. Thus the tightest test is when the constant of proportionality equals 1. There are 10 literals in the grammar of Figure 2. Thus the experimental hypothesis is that the error is at most $\frac{12}{36} = 0.33$, or 33%. This upperbound clearly holds for both cross-validation and hold-out results.

3 Discussion

This paper discusses a simple experiment involving grammars from Early Reader books. Despite the obvious analogies with the acquisition of grammars by children, the author knows of no similar previous experiments. The grammar for Book 1 was learned in under 1 minute of CPU time, and was simple and easy to comprehend. On the downside, considerable phrase structure background knowledge was required. Further progress will require the ability to automatically revise this background knowledge.

Cross-validation results (98.61%) for learning from Book 1 are at considerable variance with hold-out results (90.79%) based on Book 2. This is only to be expected due to the fact that Ladybird books provide graded material of increasing complexity. However, this demonstrates a weakness of the standard assumption within PAC learning, i.e. that all examples for training and testing are drawn from a single distribution. The indications of the pilot experiment are that a revised theoretical model is needed to cope with situations involving tutoring.

It is believed that the initial results in this paper lay the ground work for a project to test whether machine learning could be used to learn a realistic grammar at the level of competence of a five year old child. The author believes this to be a project of vital interest and importance to both Machine Learning and Natural Language research.

Acknowledgements

The author would like to thank David Haussler for influential discussions on the topic of learning from positive examples. The author's investigations with David Haussler of PAC upper-bound results for learning from positive examples will be reported elsewhere. The author is grateful to Nick Chater of the Experimental Psychology Department in Oxford for pointing out the relevant literature on language learning in children. Thanks also for useful discussions on the topics in this paper with Donald Michie, John McCarthy, Tony Hoare, David Page and Ashwin Srinivasan. This work was supported partly by the Esprit Long Term Research

Action ILP II (project 20237), EPSRC grant GR/J46623 on Experimental Application and Development of ILP, EPSRC grant GR/K57985 on Experiments with Distribution-based Machine Learning and an EPSRC Advanced Research Fellowship held by the author. The author is also a Research Fellow at Wolfson College Oxford.

References

[1] A.W. Biermann and J.A. Feldman. On the synthesis of finite-state machines from samples of their behaviour. *IEEE Transactions on Computers*, C(21):592–597, 1972.

[2] W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. Memory-based part of speech tagging. In *CLIN'95 Workshop*, Antwerp, 1995.

[3] E.M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[4] Pat Langley and Herbert Simon. Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):54–64, 1995.

[5] R.J. Mooney and M.E. Califf. Induction of first-order decision lists: Results on learning the past tense of english verbs. *Journal of Artificial Intelligence Research*, 3:1–24, 1995.

[6] S. Muggleton. Learning from positive data. In *Proceedings of the Sixth Inductive Logic Programming Workshop*, Stockholm University, 1996. (Submitted).

[7] S. Muggleton. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming*. IOS Press/Ohmsha, 1996. (To appear).

[8] Corby J. Murray W. and Russell C. *Book 1, Read with me: Let's play.* Laybird, Loughborough, UK, 1993.

[9] Corby J. Murray W. and Russell C. *Book 2, Read with me: the dragon den.* Laybird, Loughborough, UK, 1993.

A Book 1 sentences.

Here is Kate.
Here is Tom.
Here is Sam.
Sam is a dog.
Here is Kate and here is Sam.

No no.
Tom is in here and Sam is in here.
Kate likes the dog and Tom likes the dog.
Sam likes Tom and Sam likes Kate.
No Sam no.
I like Tom.
I like Kate.
I like the dog.
Here is Tom and here is Kate.
No Sam no.
Here is a shop.
The dog is in the shop.
Tom is in the toy shop.
Kate is in the toy shop.
Sam is in the toy shop.
Sam has a toy.
Sam likes the toy.
No Sam no.
Kate and Tom like the ball.
Sam likes the ball.
The dog has the ball.
The dog likes the ball.
Here is a tree.
The ball is in the tree and Tom is in the tree.
Tom has the ball.
Tom and Kate like the ball.
Sam has the ball.
No Sam no.
Sam has no toy.
Sam has the toy.
Sam likes the toy.
I like Kate and I like Tom and I like Sam.

B Book 1 random negative examples.

In the following 'EMPTY' stands for the empty sentence.

Dog.
EMPTY.
Here is Tom and like here here and Kate Kate Sam is like.
Here Tom Tom the has.
A likes Tom toy and dog Kate dog.
The dog the shop is.

Toy the and the.
Kate.
Like like the in.
Has and.
Toy I like ball dog Tom.
EMPTY.
The here ball shop dog is is ball.
Has Tom and dog shop ball has Kate likes here the Sam Kate Tom.
The here Tom Kate tree tree.
And toy is toy Sam no toy.
Here Kate in no like Tom Tom is i.
Ball Sam like the Sam toy ball Sam no.
Toy here in Kate dog toy Sam no the.
The the the a in is is the in the here Kate and has is ball Kate and Kate is I I Sam likes I is Tom no is Tom.
Sam has.
Tom here is.
Tom in is likes the has Kate is like tree.
Likes like the tree like is I the Tom Tom Sam.
The tree Kate Kate is is Kate likes.
The.
Dog.
EMPTY.
Sam and like Sam is is and likes and here the and likes.
Toy Sam like no.
Tom shop and a toy.
Like shop the Tom toy Kate the the is here shop is has like the Tom Sam here has is the.
In has Sam.
Is here I shop no shop the Sam.
Likes Sam is.
EMPTY.

C Book 1 background knowledge.

```
np(S1,S2) :- snp(S1,S2).
np(S1,S2) :- det(S1,S3), noun(S3,S2).
np(S1,S2) :- adjp(S1,S3), noun(S3,S2).
np(S1,S2) :- det(S1,S3), adjp(S3,S4), noun(S4,S2).
np(S1,S2) :- snp(S1,S3), conj(S3,S4), snp(S4,S2).
```

```
snp(S1,S2) :- noun(S1,S2).
snp(S1,S2) :- pron(S1,S2).
snp(S1,S2) :- propname(S1,S2).

det([a|S],S).
det([the|S],S).

vp(S1,S2) :- avp(S1,S2).
vp(S1,S2) :- avp(S1,S3), prep(S3,S2).

avp(S1,S2) :- tverb(S1,S2).
avp(S1,S2) :- tverb(S1,S3), adv(S3,S2).
avp(S1,S2) :- adv(S1,S3), tverb(S3,S2).

adjp(S1,S2) :- adj(S1,S2).
adjp(S1,S2) :- adv(S1,S3), adj(S3,S2).
adjp(S1,S2) :- adj(S1,S3), adj(S3,S2).

propname([tom|S],S).
propname([kate|S],S).
propname([sam|S],S).

noun([dog|S],S).
noun([shop|S],S).
noun([toy|S],S).
noun([ball|S],S).
noun([tree|S],S).
noun([here|S],S).

adj([no|S],S).
adj(S1,S2) :- noun(S1,S2).

pron([i|S],S).

tverb([like|S],S).
tverb([likes|S],S).
tverb([has|S],S).
tverb([is|S],S).

prep([in|S],S).

conj([and|S],S).
```