Journal of Machine Learning Research 12 (2011) 1425-1428 Submitted 3/11; Published 4/11

Introduction to the Special Topic on Grammar Induction, Representation of Language and Language Learning

Dorota Głowacka D.GLOWACKA@CS.UCL.AC.UK
John Shawe-Taylor JST@CS.UCL.AC.UK
Department of Computer Science
University College London
London WC1E 6BT
United Kingdom

Alexander Clark ALEXC@CS.RHUL.AC.UK
Department of Computer Science,
Royal Holloway, University of London
Egham, Surrey, TW20 0EX
United Kingdom

Colin de la Higuera CDLH@UNIV-NANTES.FR
Laboratoire LINA
University of Nantes
44322 Nantes
France

Mark Johnson MARK.JOHNSON@MQ.EDU.AU
Department of Computing
Macquarie University
Sydney NSW 2109
Australia

Editor: Lawrence Saul

Abstract

Grammar induction refers to the process of learning grammars and languages from data; this finds a variety of applications in syntactic pattern recognition, the modeling of natural language acquisition, data mining and machine translation. This special topic contains several papers presenting some of recent developments in the area of grammar induction and language learning, as applied to various problems in Natural Language Processing, including supervised and unsupervised parsing and statistical machine translation.

Keywords: machine translation, Bayesian inference, grammar induction, natural language parsing

1. Introduction

Grammar induction was the subject of an intense study in the early days of Computational Learning Theory, with the theory of query learning (Angluin, 1988) largely developing out of this research. More recently the study of new methods of representing language and grammars through complex kernels and probabilistic modelling together with algorithms such as structured output learning has enabled machine learning methods to be applied successfully to a range of language related tasks from simple topic classification through parts of speech tagging to statistical machine translation. These methods typically rely on more fluid structures than those derived from formal grammars and yet are able to compete favourably with classical grammatical approaches that require significant input from domain experts, often in the form of annotated data and hand-coded rules.

2. JMLR Special Topic

This special topic arose from a NIPS 2009 workshop on ”Grammar Induction, Representation of Language and Language Learning” held at the Whistler Resort, Vancouver, Canada. Contributions to the special topic were also open to researchers who had not presented their work at the workshop. We received thirteen submissions and after considering the reviews for each submission, we selected five papers to be included in this special topic.

Probabilistic grammars offer great flexibility in modeling discrete sequential data like natural language text. Recently, there has been an increased interest in using probabilistic grammars in the Bayesian setting, focusing mostly on the use of a Dirichlet prior. Cohen and Smith (2010) propose a family of logistic normal distributions as an alternative to the Dirichlet prior. A variational inference algorithm for estimating the parameters of the probabilistic grammar provides a fast, parallelizable, and deterministic alternative to MCMC methods to approximate the posterior over derivations and grammar parameters. Experiments with dependency grammar induction on six different languages demonstrate performance improvements with the new priors. The experiments include a novel promising setting, in which syntactic trees are inferred in a bilingual setting that uses multilingual, non-parallel corpora. Notably, the proposed approach tends to generalize better to longer sentences, despite learning on short sentences.

Despite decades of research, inducing a grammar from text has proven to be a notoriously challenging learning task. The majority of existing work on grammar induction has favoured model simplicity (and thus learnability) over representational capacity by using context free grammars and first order dependency grammars, which are not sufficiently expressive to model many common linguistic constructions. Cohn, Blunsom, and Goldwater (2010) propose a novel compromise by inferring a Probabilistic Tree Substitution Grammar (PTSG), a formalism which allows for arbitrarily large tree fragments and thereby better represents complex linguistic structures. A PTSG is an extension to the Probabilistic Context Free Grammar (PCFG) in which nonterminals can rewrite as entire tree fragments (elementary trees), not just immediate children. These large fragments can be used to encode non-local context, such as argument frames, gender agreement and idioms. The model’s complexity is reduced by employing a Bayesian non-parametric prior which biases the model towards a sparse grammar with shallow productions. The experimental results demonstrate the model’s efficacy on supervised phrase-structure parsing, where a latent segmentation of the training treebank is induced, and on unsupervised dependency grammar induction. In both cases the model uncovers interesting latent linguistic structures while producing competitive results.

Henderson and Titov (2010) propose a new class of graphical models for structured prediction problems called incremental sigmoid belief networks (ISBNs) and apply it to natural language grammar learning. ISBNs make decoding possible because inference with partial output structures does not require summing over the unboundedly many compatible model structures, due to their directed edges and incrementally specified model structure. ISBNs are particularly applicable to natural language parsing, where learning the domain’s complex statistical dependencies benefits from large numbers of latent variables. Exact inference in ISBNs is not tractable, but two efficient approximations are proposed: a coarse mean-field approximation and a feed-forward neural network approximation. Experimental results show that these models achieve accuracy competitive with the state-of-the-art.

Machine translation is a challenging problem in artificial intelligence. Natural languages are characterised by large variabilities of expressions, exceptions to grammatical rules and context dependent changes, making automatic translation a very difficult task. While early work in machine translation was dominated by rule based approaches (Bennett and Slocum, 1985), the availability of large corpora has paved the way for statistical methods to be applied. Ni, Saunders, Szedmak, and Niranjan (2011) propose a distance phrase reordering model (DPR) for statistical machine translation, where the aim is to learn the grammatical rules and context dependent changes using a phrase reordering classification framework. Techniques are compared and evaluated on a Chinese-English corpus, a language pair known for the high reordering characteristics which cannot be adequately captured with current models. In the reordering classification task, the method significantly outperforms the baseline against which it was tested, and further, when integrated as a component of the state-of-the-art machine translation system, MOSES, it achieves improvement in translation results.

Gillenwater, Ganchev, Graça, Pereira, and Taskar (2011) present a new method for unsupervised learning of dependency parsers. In contrast with previous approaches that impose a sparsity bias on the model parameters using discounting Dirichlet distributions, the proposed technique imposes a sparsity bias on the model posteriors. This is done by using the posterior regularization (PR) framework (Graça et al., 2007) with constraints that favor posterior distributions that have a small number of unique parent-child relations. In experiments with 12 different languages, the proposed method achieves significant gains in directed accuracy over the standard expectation maximization (EM) baseline for 9 of the languages, while for 8 out of 12 languages, the new technique outperforms models based on standard Bayesian sparsity-inducing parameter priors.

3. Concluding Remarks

We feel these papers provide a useful snapshot of the current state-of-the-art techniques being employed by researchers in the fields of grammar induction, language parsing, machine translation and related areas.

Acknowledgments

We are much indebted to our authors and reviewers, who put a tremendous amount of effort and dedication to make this issue a reality. We are also thankful to the JMLR editorial team for their valuable support.

References

D. Angluin. Queries and concept learning. *Machine Learning*, 2:319 – 342, 1988.

W. S. Bennett and J. Slocum. The lrc machine translation system. *Computational Linguistics*, 11: 111 – 121, 1985.

S. B. Cohen and N. A. Smith. Covariance in unsupervised learning of probabilistic grammars. *Journal of Machine Learning Research*, 11:3017 – 3051, 2010.

T. Cohn, P. Blunsom, and S. Goldwater. Inducing tree-substitution grammars. *Journal of Machine Learning Research*, 11:3053 – 3096, 2010.

J. Gillenwater, K. Ganchev, J. Graça, F. Pereira, and B. Taskar. Posterior sparsity in unsupervised dependency parsing. *Journal of Machine Learning Research*, 11, 2011.

J. Graça, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, Cambridge, MA, 2007. MIT Press.

J. Henderson and I. Titov. Incremental sigmoid belief networks for grammar learning. *Journal of Machine Learning Research*, 11:3541 – 3570, 2010.

Y. Ni, C. Saunders, S. Szedmak, and M. Niranjan. Exploitation of machine learning techniques in modelling phrase movements for machine translation. *Journal of Machine Learning Research*, 12:1 – 30, 2011.