Machine Learning Algorithms for Natural Language Processing and Information Extraction

Michael Collins

1 Introduction

Large databases of text are becoming increasingly prevalent, obvious examples being web data or digital libraries. Algorithms which recover structure underlying this data are therefore becoming increasingly important. The goal of this project is to develop machine learning algorithms for a variety of natural language processing and information extraction tasks. The eventual aim is to employ these algorithms in aiding users to search, browse, or extract information from large repositories of text.

We plan to focus on two related research areas. The first is development of approaches to novel information extraction tasks. The second is to continue to develop machine learning algorithms for NLP and information extraction.

2 Proposed Work

Research on Novel Information Extraction Tasks: Much of the published work in statistical or machine learning approaches to information extraction has been on tasks that involve fairly "shallow" processing. One example is named-entity extraction (identifying people, locations, organizations etc.), which has been studied extensively, with quite successful results. In contrast there has been relatively little published research on learning approaches which go beyond this level of analysis, for example to extract relationships between entities (although see [Miller et. al 2000] for one learning approach to this problem). Our plan is to study information extraction tasks which go beyond named-entity recognition, to build "deeper" analyses of the documents in a corpus. This will very likely involve statistical approaches to parsing which have been developed in previous work [Collins 1999].

Development of New Learning Algorithms: There are a number of reasons for developing new algorithms for learning approaches to NLP and information extraction. There is a need for methods which give improved accuracy on tasks, and which are more flexible in the representations they can utilize. There is a need for algorithms which are more computationally efficient in both training and in their application to new examples. Crucially, because supervised training examples are often expensive to collect, there is an acute need for methods which require less supervised data, or even algorithms which depend almost entirely on unlabeled data. Novel information extraction tasks (the first part of this proposal) will undoubtedly require new methods. Finally, there is a need for a greater theoretical understanding of the problems that arise in NLP and information extraction tasks. Much of the work in learning theory has focused on classification and regression problems – the problems we are considering are somewhat different, in that they involve a mapping from one discrete set to another (for example mappings from strings to strings, or from strings to trees).

3 Technical Approach

Many of the tools for information extraction and statistical NLP are based on weighted or stochastic variants of finite-state automata and context-free grammars. We intend to continue research on these methods. More generally, we intend to investigate representations and algorithms used in the graphical models community – these may lead to richer representations of more complex information extraction tasks, and also to the use of algorithms such as loopy belief propagation and variational methods [Jaakkola and Jordan 1999] which are relatively novel to the NLP and IE communities. For the use of unlabeled data, methods might depend on the EM algorithm, or on the CoTraining approach [Blum and Mitchell 98]. We intend to investigate further the use of improved representations, either through "global" features [Collins 2000], or kernel methods over discrete structures (e.g., [Collins and Duffy 2002]). Finally, we intend to continue research on discriminative training methods based on boosting, perceptron and support vector machine approaches originally developed for classification tasks [Collins 2000, Collins 2002, Collins and Duffy 2002].

References

[Blum and Mitchell 98] Blum, A., and Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 1998 Conference on Computational Learning Theory.*

[Collins 1999] Collins, M. 1999. Head-Driven Statistical Models for Natural Language Parsing. *PhD Thesis, University of Pennsylvania.*

[Collins 2000] Collins, M. (2000). Discriminative Reranking for Natural Language Parsing. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000).*

[Collins and Duffy 2002] Collins, M., and Duffy, N. (2002). New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *Proceedings of ACL 2002.*

[Collins 2002] Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with the Perceptron Algorithm. In *Proceedings of EMNLP 2002.*

[Jaakkola and Jordan 1999] Jaakkola, T., and Jordan, M. 1999. Variational probabilistic inference and the qmr-dt database. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

[Miller et. al 2000] Miller, S., Fox, H., Ramshaw, L., and Weischedel, R. 2000. A Novel Use of Statistical Parsing to Extract Information from Text. In *Proceedings of ANLP 2000.*