Theory and Algorithms for Information Extraction and Classification in Textual Data Mining

Tianhao Wu
CSE Department, Lehigh University
May 27, 2003

1. Introduction

Regular expressions can be used as patterns to extract features from semi-structured and narrative text [8]. For example, in police reports a suspect's height might be recorded as "{CD} feet {CD} inches tall", where {CD} is the part of speech tag for a numeric value. The result in [1] shows us that regular expressions could have higher performance than explicit expressions in some applications such as Posting Act Tagging. Although much work has been done in the field of information extraction, relatively little has focused on the automatic discovery of regular expressions. Therefore, my Ph.D. research will focus on the automatic generation of reduced regular expressions (RREs) (defined in [8]) used in Information Extraction (IE).

The reduced regular expressions learned can be directly used to extract features from free text, or they can be used to fill in templates in Eric Brill's Transformation-Based Learning (TBL) [2] frameworks. The original templates in TBL are explicit expressions, which are weaker than reduced regular expressions. I propose an innovative enhancement to TBL termed "Error-Driven Boolean-Logic-Rule-Based Learning" (BLogRBL) [9], which is strictly more powerful than TBL [2]. Similar to Brill's method, rules are automatically derived from templates during learning. It differs from Brill's technique in that rules take the form of complex expressions of combinational logic. Therefore, my final contribution in my PhD thesis will be a framework that combines regular expression discovery with BLogRBL.

A necessary component of this research is a study of various biases inherent in the use of reduced regular expressions in IE. The purpose of this work is to determine the language biases, search biases, and overfitting biases in the RRE discovery and BLogRBL algorithms.

2. Related Work

There are two different types of related work discussed in this section. One is work related to IE using regular expressions. The other is variations of Eric Brill's Transformation-Based Learning (TBL). In this section, I will first describe two efforts using regular expressions in IE. Following this, I will introduce work employing variations of TBL.

Stephen Soderland developed a supervised learning algorithm, WHISK [3], which uses regular expressions as patterns to extract features from semi-structured and narrative text. In each iteration of the learning process, WHISK requires that a human expert label specific features in instances and then generates rules based on these labels. WHISK uses segments such as clauses, sentences, or sentence fragments as its instances. A crucial difference between WHISK and our approach is that WHISK requires the user to identify the precise location of features for labeling while our approach requires only that instances be labeled. As noted this represents a significant reduction in the effort required to develop a training set.

Eric Brill [4] applied his transformation-based learning (TBL) framework to learn reduced regular expressions for correction of grammatical errors in text. Although Brill does not perform explicit information extraction, the correction process involves identifying grammatical errors. There are two major differences between Brill's approach and ours. First, the reduced regular expressions generated by Brill do not include the logical "OR" operator. We have found that the "OR" operator is necessary to achieve high accuracies in information extraction. Secondly, like the aforementioned work by Soderland, Brill's approach requires intensive feature-specific labeling to create the ground truth used in TBL.

Several enhancements and modifications have been made since Brill introduced TBL. Here I briefly review a few examples of this work.

K. Samuel and K. Vijay-Shanker developed a Monte Carlo version of TBL [5], which randomly selects a subset of rules in each learning iteration to apply. Since not all rules are applied, the learning process is faster. The authors also used a 'committee method' to increase the overall performance of TBL.

Lidia Mangu [6] used statistical significance to find the best 'stop rule' rather than the lowest error rate in TBL. Since this approach ignores some rules in learning iterations, the learning process is faster. At the same time, it also loses some accuracy.

David Palmer [7] used a balanced F_β-measure instead of an error rate as the scoring function of TBL for Chinese word segmentation. The balanced F_β-measure handled precision and recall equally in weight. Precision and recall are more important than error rate for some NLP problems [7]. Therefore, we also use the balanced F_β-measure in BLogRBL.

3. Proposed Research

My proposed Ph.D. research will be in three major phases, and each phase has several steps. The first task of my research is to create a semi-supervised learning algorithm that can generate reduced regular expressions from a small training dataset. The algorithm is named "Reduced Regular Expression Discovery" (RRE Discovery) algorithm. Next, I will extend TBL to BLogRBL. In this phase, I will provide comparison of TBL and BLogRBL in theory and empirically in application experiments. Lastly, I will combine these two approaches together based on the diagram in Figure 1. Following this, I will study biases of the RRE Discovery algorithm and BLogRBL. The details of each phase are considered in what follows.

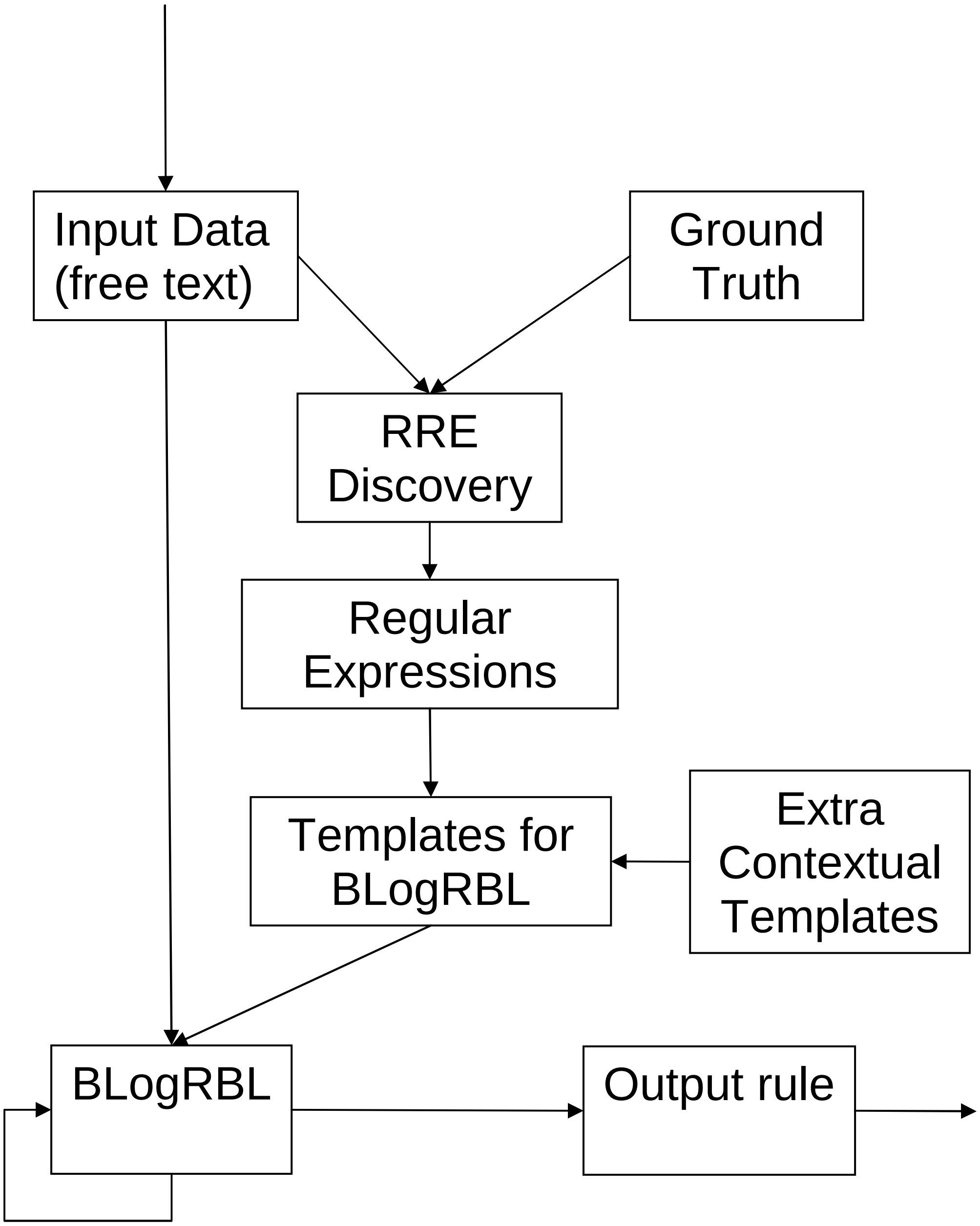

Figure 1: BLogRBL with RRE Discovery as the template generator

3.1 Reduced Regular Expression Discovery

We have developed a covering algorithm that discovers reduced regular expressions with "AND", "OR", "NOT", and "Optional" operators. The formal definition of a reduced regular expression is given in [8]. The system has been tested successfully on ten oft-used features present in Fairfax County, Virginia police incident reports[1]. The algorithm has also been applied to extract features from US patents that identify the

[1] www.co.fairfax.va.us/ps/police/reports

problem that a given patent addresses. We term these features Problem Solved Identifiers (PSIs). A summary of this application appears in [10]. Most of the work in this phase has been completed, and is reported in [8] and [10]. A prototype of the algorithm has been implemented in Perl [11] [12]. The remaining research in this phase is to refine the current "RRE Discovery" algorithm to improve its performance and reduce the amount of training data required to discover an RRE.

3.2 Error-Driven Boolean-Logic-Rule-Based Learning

In the second phase of my research, I will continue to explore BLogRBL. We have published a technical report that describes BLogRBL in [9]. In the report, we compare BLogRBL and TBL. We prove that BLogRBL is strictly more powerful than TBL. One of the tasks ahead is to identify applications in which BLogRBL has better performance than TBL. One possible application is the detection of PSIs since this is a two-class classification problem, which is suitable for both TBL and BLogRBL. Therefore, the performance of these two approaches can be readily compared.

3.3 The Combination of RRE Discovery and BLogRBL

I plan to use the framework depicted in Figure 1 to combine BLogRBL and the RRE Discovery algorithm. Input data such as annotated text and the ground truth will be fed to the RRE Discovery algorithm. Templates for BLogRBL can be generated by generalizing reduced regular expressions discovered by the RRE Discovery algorithm. Other non-regular-expression templates (extra contextual templates) will also be supported in this framework. Even though the reduced regular expressions include contextual information, we expect that some contextual templates may be more readily described using a representation other than RREs. For example, the offset of a feature might be a template for BLogRBL that is more easily represented as a standard TBL/BLogRBL contextual template as opposed to an RRE.

Input data and templates will be processed by BLogRBL, which will result in a single Boolean rule that can used to extract information from previously unseen data.

3.4 Biases in RRE Discovery and BLogRBL

The final step in my Ph.D. research is to analyze the algorithms' biases. All algorithms have various biases [13]. The RRE Discovery algorithm is no exception. For instance, a language bias of the algorithm is that it uses reduced regular expressions that support a subset of regular expressions such as (rs), where 'r' and 's' are RREs. This algorithm does not support other subsets of regular expressions such as "α*", where 'α' is a RRE and '*' indicates that the character immediately to its left may occur any number of times, including zero. An example of search bias is that the algorithm attempts to identify words or part-of-speech tags that occur most often in the ground truth. I propose to fully characterize the various biases present in both our RRE discovery algorithm as well as in

BLogRBL. Based on the algorithms' biases, I will ascertain which applications are suitable for the algorithms, and which are not.

In this section, I have briefly described the proposed phases in my Ph.D. research. In each phase, concrete steps have also been discussed. Two major research directions, IE using the RRE Discovery algorithm and BLogRBL, will be explored. My goal is to combine these approaches to form a new algorithm for information extraction and classification that is strictly more powerful than TBL.

References

[1] Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler and William M. Pottenger. *Posting Act Tagging Using Transformation-Based Learning*. In the Proceedings of the Workshop on Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining (ICDM'02). December 2002.

[2] Brill, Eric. Transformation-Based Error Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics 21(94): 543-566. 1995.

[3] S. Soderland. *Learning Information Extraction Rules for Semi-structured and Free Text.* Machine Learning, 34(1-3):233-272, (1999).

[4] Eric Brill. *Pattern-Based Disambiguation for Natural Language Processing*. Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, (2000).

[5] K. Samuel and K. Vijay-Shanker. Dialogue Act Tagging with Transformation-Based Learning. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL-COLING), pp. 1150-1156, 1998.

[6] L. Mangu and E. Brill (1997). *Automatic Rule Acquisition for Spelling Correction.* In Proc. of the Fourteenth International Conference on Machine Learning, ICML'97, Nashville, Tennessee.

[7] David Palmer. 1997. *A Trainable Rule-Based Algorithm for Word Segmentation.* Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97), Madrid, 1997.

[8] Tianhao Wu, and William M. Pottenger. *A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data*. The seventh Pacific-Asia conference on Knowledge Discovery and Data Mining (PAKDD), April 2003.

[9] Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler and William M. Pottenger. *Error-Driven Boolean-Logic-Rule-Based Learning for Mining Chat-room Conversations.* Lehigh University CSE department technical reports. LU-CSE-02-008. 2002.

[10] Tianhao Wu and William M. Pottenger. "*A Supervised Learning Algorithm for Information Extraction from Textual Data"*. In the proceeding of the workshop on Text Mining, Third SIAM International Conference on Data Mining, San Francisco, May (2003).

[11] E.F. Friedl, "Mastering regular expressions", O'Reilly & Associates, Inc., Sebastopol, CA, 1997.

[12] Larry Wall, Tom Christiansen and John Orwarnt. "Programming Perl". , O'Reilly & Associates, Inc.

[13] Alex A. Freitas and Simon H. Lavington. "Mining Very Large Databases with Parallel Processing". Kluwer Academic Publishers, 1998.