Dissertation

GROPUS – an adaptive rule-based algorithm for information extraction

---

Peter Siniakov

Freie Universität Berlin
Fachbereich Mathematik und Informatik

Supervisors: Prof. Dr.-Ing. Heinz F. Schweppe
Prof. Dr. Ulf Leser

Eingereicht am 29. November 2007
Disputation am 11. April 2008

Abstract

Internet, mass media, scientific literature are the source of huge, continuously growing amount of information that is comprised by natural language texts and stored in digital form. This information can hardly be immediately accessed and processed by computers while human access is often connected with a time-consuming search. Extracting and storing it in a formal representation (e.g. in form of relations in databases) allows efficient querying and easy administration of the extracted data. Moreover, information stored and queried in a canonical way can be processed and interpreted by computers without human interaction; it can serve for establishing ontologies, creation of knowledge bases and data analysis. The area of IE comprises techniques, algorithms and methods performing two important tasks: finding (identifying) the desired, relevant data in natural language texts and storing it in a structured representation suitable for automatic processing.

First IE systems relied on domain-specific extraction rules written by a domain expert requiring large human effort and lacking portability to other domains. To compensate the insufficiencies of the classical rule-based approach human effort should be adequately replaced by a learning component. The main goal of my dissertation has been the development of an adaptive, rule-based algorithm for IE that autonomously learns the extraction rules. The algorithm is based on induction learning deriving general extraction rules from a set of sample extractions annotated by a human in a training corpus. Requiring only an annotated training corpus and no additional resources the approach is portable to different application domains and even languages (in the dissertation its effectiveness for English and German text corpora has been examined).

The extraction rules incorporate linguistic patterns that capture typical expression forms of extracted information in a given text corpus. We introduce a higher-order formal pattern specification language that supports regular expressions, permutation, negation and hierarchical XML structures significantly extending common pattern models. Linguistic patterns are not restricted to a fix context window, but encode whole sentences as primary semantic units of natural language. The proposed pattern language is powerful and expressive enough to capture non-trivial kinds of phrases and sentences containing relevant information.

The linguistic patterns are matched with linguistically preprocessed texts that have a valid XML markup. Regarding linguistic patterns as XML queries we reduce the problem of IE to XML query evaluation. Having developed formal semantics and an efficient query evaluation algorithm for the pattern language we create a new XML query language, which is especially suitable for querying XML annotated texts.

As a part of semantic text preprocessing we propose a new method for determination of synonymy. We construct a lexical graph connecting lexical items in a way corresponding to the sentence structure building an implicit context representation. We demonstrate that the synonymy metric based on the length of paths between two lexical items, number and specificity of shared neighbors achieves satisfactory results evaluating it on a test corpus of 200 German synonyms. Identified synonyms are used for abstraction of lexical items during the rule induction.

Beginning with the rules generated from training instances, which were extracted by the human, rules are generalized to account for different kinds of information expression in the texts. The generalization of rules is formally specified and involves beside rule merging abstraction of single rules and substitution of extracted parts in context of different rules. For establishing a similarity measure for extraction rules and rule merging an algorithm for determination of optimal alignment of two sequences with minimum runtime (which is an extension of the LCS problem) has been designed and its correctness proved. To achieve a gradual generalization of extraction rules the rule learning algorithm includes validation of induced rules and rule correction.

We demonstrate the effectiveness of our approach comparing its performance with other state of the art approaches achieving comparable or even best results depending on the kind of texts and assess its potential comparing its results with the human performance. Based on varying performance of different approaches on different corpora conclusions about the efficiency of statistical and rule-based approaches for different kinds of text are made. The quantitative investigation is supplemented by the analysis what factors influence the extraction quality, what are the sources of errors etc. Finally, we draw a conclusion in what conditions application of IE in general is expedient, what kinds of text can be managed and characterize the range of environments where the presented approach can be usefully utilized.

Acknowledgements

The research that culminated in this dissertation has been initiated by the FEx project at the Database & Information Systems Group at the CS department of Freie Universität Berlin. I owe a great debt of gratitude to the chair of the group and my first supervisor Prof. Heinz Schweppe for his inspiration and friendship. Having always the open door for questions and exchange of ideas he created a productive environment in that many fruitful discussions took place and that helped to broaden my scientific horizon.

My warm thanks go to the other members of the FEx project, Heiko Kahmann – for his verve and enthusiasm in our joint work, Christian Siefkes – for many controversial, yet constructive discussions and to both of them for the great time we had together.

I am very grateful to my second supervisor, Prof. Ulf Leser for his critical view and valuable advises and comments.

I have very appreciated a pleasant working atmosphere established by the members of the Database Group who have always been a lighthearted and cohesive collective.

Furthermore I would like to thank Prof. Leake from the CS department of Indiana University for sparking my interest in reasoning and semantic text processing by his original and captivating courses.

And finally I want to deeply thank my grandfather, Aron Siniakov, who has always been a wise mentor and a role model for me.

This research was supported by the NAFöG Scholarship of the federal state Berlin.

Contents

I Information Extraction: Problems and Perspectives **11**

1 Introduction **12**

1.1 Information Extraction 12

1.2 Related Research Areas 13

1.3 The Problem of IE 14

1.4 Why is the problem of IE not solved? 15

1.5 Outline of this Work 16

2 Related Work - Adaptive Approaches to IE **18**

2.1 Relevant Machine Learning Techniques 19

2.1.1 Classification 20

2.1.2 Determining the success of learning 20

2.1.3 Rule Learning 21

2.2 Rule-Based Approaches 21

2.2.1 Automatic pattern and template creation 21

2.2.2 Covering algorithms 23

2.2.3 Relational Learners 24

2.3 Knowledge-based and Statistical Approaches 26

2.4 Deficiencies of state of the art approaches 28

3 Approaching the Problem **30**

3.1 Types of Information in the Natural Language Texts 30

3.2 Choosing an Appropriate Approach 31

3.3 Goals and Requirements 34

II Inductive Learning of Extraction Rules **36**

4 Rule-Based Approach to Information Extraction **37**

4.1 Overview of the system 37

4.2 Text corpus and target structure 42

4.2.1 Characteristics of textual input 42

4.2.2 Accepted text formats 43

4.2.3 Human annotations 43

4.2.4 Target Structure 44

5 Preprocessing of Text Corpus **46**

5.1 Linguistic preprocessing 46

5.2 Semantic Preprocessing: Recognition of Synonyms 49

5.2.1 Approaches to Recognition of Synonymy 50

5.2.2 Construction of the lexical graph 51

5.2.3 Identification of synonyms 53

5.2.4 Experiments with Synonym Recognition 55

5.2.5 Utilization of Synonym Recognition for Abstraction of Extraction Patterns 58

6 Pattern Unification by Querying XML: A Pattern Based XML Query Language **60**

6.1 Pattern Specification Language in the Role of XML Query Language 61

6.2 Existing XML Query Languages 62

6.3 Querying by pattern matching 63

6.3.1 Extended sequence semantics 64

6.3.2 Specification of XML nodes 67

6.3.3 Backtracking patterns and variables 67

6.3.4 Negation 69

6.4 Unification Algorithm 70

6.4.1 Unification of Negation Pattern 71

6.4.2 Handling Backtracking and Assignment Patterns 71

6.4.3 Assessment of Time Complexity 72

6.5 Summary 74

7 Induction of Extraction Rules **75**

7.1 Generation of Initial Rules 75

7.1.1 Localization of Extracted Fragments in the Training Documents 75

7.1.2 Choosing the Appropriate Context 76

7.1.3 Translation of Extractions and their Context in Pattern Language 77

7.1.4 Encoding of Extractions 78
7.2 Rule Similarity 80
7.2.1 Mapping Hierarchies to Sequences for Comparison 81
7.2.2 Algorithm for Comparison of Sequence Similarity 83
7.2.3 Rule Similarity Measure 87

8 Rule Generalization and Correction 89
8.1 Rule Merging 89
8.1.1 Merging Lexical Strings, POS and XML Patterns 91
8.1.2 Merging Sequences 92
8.1.3 Generalizing Differences by Backtracking Patterns 92
8.1.4 Recursive Usage of Backtracking Patterns 93
8.1.5 Merging of Complete Extraction Rules 94
8.2 Rule Abstraction 95
8.2.1 Relaxation of Context 96
8.2.2 Abstracting Function 96
8.3 Substitution Heuristic 96
8.4 Rule Correction 97
8.4.1 Types of Errors 98
8.4.2 Rule Correction Algorithm 99
8.4.3 Limitations of Rule Correction 101

9 Learning Algorithm and Application 102
9.1 Selection of Extraction Rules for Generalization 102
9.1.1 Rule Inclusion and Abstraction Degree 103
9.1.2 Controlling the Rule Generalization 104
9.1.3 Runtime of the Rule Induction 106
9.1.4 Utilization of Rule Abstraction and Substitution 107
9.2 Validation of Induced Rules 108
9.2.1 The Purpose of Validation 108
9.2.2 Rule and Attribute Precision Thresholds 109
9.2.3 Local Attribute Precision Thresholds 112
9.2.4 Covering Validation Setup 112
9.3 Termination of Rule Induction and Application 113
9.3.1 Termination 113
9.3.2 Application of Learned Extraction Rules 114

III Evaluation 115

10 Introduction to the Empirical Investigation 117

10.1 Test Corpora . . . 117
10.1.1 Seminar Announcement Corpus . . . 119
10.1.2 Bosnian Corpus . . . 119
10.1.3 MUC corpus . . . 120
10.1.4 Comparison of Corpora . . . 122
10.2 Investigated Questions . . . 122
10.3 Evaluation Methodology . . . 124
10.3.1 Quantitative Metrics . . . 124
10.3.2 Experimental Setup and Determination of Total Values . 125
10.3.3 Evaluation Modes . . . 126

11 Extraction Results and Comparison with Other IE Approaches and Human Performance 127
11.1 Experiments with Bosnian Corpus . . . 127
11.1.1 Behavior of GROPUS for single attributes . . . 128
11.1.2 Comparison with TIE on Original and Preclassified Corpus 129
11.2 Experiments with Seminar Announcement Corpus . . . 131
11.2.1 Analysis of the Quality of Extractions from Semistructured Texts . . . 131
11.2.2 Discussion of Common Errors . . . 133
11.2.3 Comparison with Other State of the Art IE Systems . . . 135
11.3 Experiments with MUC Corpus . . . 136
11.3.1 Discussion of Extraction Results Achieved by GROPUS . 137
11.3.2 Comparison with Statistical Systems TIE and ELIE . . . 139
11.4 Interpretation of Results in the Face of Human Performance . . 140
11.4.1 Peculiarities of the Training Data for IE Based on Supervised Learning . . . 141
11.4.2 Evaluation of Human Results . . . 141
11.5 Runtime Comparison . . . 143

12 What Influences Extraction Quality? 145
12.1 Dependency of Extraction Goodness on the Size of the Training Corpus . . . 145
12.2 Text Classification as a Preparatory Step for IE . . . 149
12.2.1 Performance on the Non-classified and Manually Classified Corpus . . . 150
12.2.2 Impact of Automatic Text Classification . . . 150
12.3 Influence of the Complexity of Attribute Values on the Extraction Quality . . . 151
12.4 Component Relevance Study . . . 152
12.4.1 Effect of Single Components on the Extraction Quality . 153

12.4.2 Evaluation of Validation Strategies 156

13 Conclusion **160**

13.1 Summary of our Approach and Contributions 160

13.2 Discussion of Results 161

13.2.1 Performance of GROPUS and other IE approaches 161

13.2.2 Influencing Factors for the Success of IE 164

13.2.3 Utility of Internal Components 165

13.2.4 In what Environments can IE be usefully employed? . . . 166

13.3 Open Problems 166

13.4 IE in Context of Knowledge Management Systems 167

13.5 Final Remarks 168

A Definition of the Pattern Language **176**

B Pattern Matching Algorithm for Selected Patterns **178**

C Algebraic Transformations for Recursive Backtracking Patterns **180**

D Anhang gemäß Promotionsordnung **181**