Dissertation (Ph.D. Thesis)

An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models

Christian Siefkes

Disputationen:
16th February 2007

Primary Supervisor: Prof. Dr. Heinz F. Schweppe
Database and Information Systems Group
Fachbereich Mathematik und Informatik
Freie Universität Berlin

Supervisors:

Prof. Dr. Heinz F. Schweppe
Database and Information Systems Group
Institute for Computer Science
Freie Universität Berlin

Prof. Dr. Bernhard Thalheim
Systems for Information Management
Institute of Computer Science and Applied Mathematics
Christian-Albrechts-Universität zu Kiel

This thesis has been supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316).

Dedicated to
the memory of my parents,

Uta Siefkes
(1940–2002)

and

Harm Siefkes
(1936–1989)

Abstract

Most of the information stored in digital form is hidden in natural language (NL) texts. While *information retrieval* (IR) helps to locate documents which might contain the facts needed, there is no way to answer queries. The purpose of *information extraction* (IE) is to find desired pieces of information in NL texts and store them in a form that is suitable for automatic querying and processing.

The goal of this thesis has been the development and evaluation of a trainable statistical IE approach. This approach introduces new functionality not supported by current IE systems, such as support for *incremental training* to reduce the human training effort by allowing a more interactive workflow.

The IE system introduced in this thesis is designed as a generic framework for statistical classification-based information extraction that allows modifying and exchanging all core components (such as classification algorithm, context representations, tagging strategies) independently of each other. The thesis includes a systematic analysis of switching one such component (the tagging strategies).

Several new sources of information are explored for improving extraction quality. Especially we introduce rich tree-based context representations that combine document structure and generic XML markup with more conventional linguistic and semantic sources of information. Preparing these rich context representations makes it necessary to unify various and partially conflicting sources of information (such as structural markup and linguistic annotations) in XML-style trees. For this purpose, we develop a merging algorithm that can repair nesting errors and related problems in XML-like input.

As the core of the classification-based IE approach, we introduce a generic classification algorithm (Winnow+OSB) that combines online learning with novel feature combination techniques. We show that this algorithm is not only suitable for information extraction, but also for other tasks such as text classification. Among other good results, the classifier was found to be one of the two best filters submitted for the 2005 Spam Filtering Task of the *Text REtrieval Conference (TREC)*.

The thesis includes a detailed evaluation of the resulting IE which shows that the results reached by our system are better than or competitive with those of other state-of-the-art IE systems. The evaluation includes an ablation study that measures the influence of various factors on the overall results and finds that all of them contribute to the good results of our system. It also includes an analysis of the utility of interactive incremental training that confirms that this newly introduced training regimen can be very helpful for reducing the human training effort. The quantitative evaluation is complemented with an analysis of the kinds of mistakes made during extraction and their likely causes that allows a better understanding of where and how we can expect further improvements in information extraction quality to be made and which limits might exist for information extraction systems in general.

Wir sehen ein kompliziertes Netz von Ähnlichkeiten, die einander übergreifen und kreuzen. Ähnlichkeiten im Großen und Kleinen.

Ich kann diese Ähnlichkeiten nicht besser charakterisieren als durch das Wort „Familienähnlichkeiten“; denn so übergreifen und kreuzen sich die verschiedenen Ähnlichkeiten, die zwischen den Gliedern einer Familie bestehen: Wuchs, Gesichtszüge, Augenfarbe, Gang, Temperament, etc. etc. – Und ich werde sagen: die „Spiele“ bilden eine Familie. [. . .]

Wie würden wir denn jemandem erklären, was ein Spiel ist? Ich glaube, wir werden ihm Spiele *beschreiben, und wir könnten der Beschreibung hinzufügen: „das,* und Ähnliches, *nennt man ‚Spiele'“. Und wissen wir selbst denn mehr? Können wir etwa nur dem Anderen nicht genau sagen, was ein Spiel ist? – Aber das ist nicht Unwissenheit. Wir kennen die Grenzen nicht, weil keine gezogen sind.*

— Ludwig Wittgenstein, Philosophische Untersuchungen

Leeloo: *Hello.*

Korben Dallas: *Oh, so you speak English now.*

Leeloo: *Yes. I learned.*

— The Fifth Element (1997)

Contents

1 Introduction **9**
1.1 Motivation and Goals 9
1.2 Contributions 10
1.3 Outline of this Work 11
1.4 Acknowledgments 11

I The Field of Information Extraction **13**

2 Information Extraction **15**
2.1 Information Retrieval, Text Mining, and Other Related Areas 15
2.2 Overview and Classification of Approaches 16

3 Architecture and Workflow **19**
3.1 Tasks to Handle 19
3.2 Architecture of a Typical IE System 21
3.3 Active Learning and Incremental Learning 25
3.4 Workflow 25

4 Statistical Approaches **29**
4.1 Probabilistic Semantic Parsing 29
4.2 Hidden Markov Models 30
4.3 Maximum Entropy Markov Models and Conditional Random Fields . 31
4.4 Token Classification 32
4.5 Fragment Classification and Bayesian Networks 33

5 Non-Statistical Approaches **35**
5.1 Covering Algorithms 35
5.2 Relational Rule Learners 37
5.3 Wrapper Induction 38
5.4 Hybrid Approaches 40
5.5 Knowledge-based Approaches 40

6 Comparison of Existing Approaches **43**
6.1 Types of Tasks Handled 43
6.2 Types of Texts Handled 44
6.3 Considered Features 44
6.4 Tagging Requirements and Learning Characteristics 45

II Analysis **47**

7 Aims and Requirements **49**

7.1 Aims of Our Approach 49

7.2 Further Requirements 51

7.3 Chosen Approach 53

7.4 Non-Goals 54

8 Assumptions **55**

8.1 Novel Assumptions 55

8.2 General Assumptions 55

8.3 Suitability of Tasks 56

9 Target Schemas and Input/Output Models **59**

9.1 Target Schemas 59

9.2 Formats for Input Texts 60

9.3 Input Formats for Answer Keys 62

9.4 Serialization of Extracted Attribute Values 65

III Algorithms and Models **67**

10 Modeling Information Extraction as a Classification Task **69**

10.1 Idea and Concept 69

10.2 Tagging Strategies 70

11 Classification Algorithm and Feature Combination Techniques **73**

11.1 The Winnow Classification Algorithm 73

11.2 Feature Combination Techniques 76

11.3 Alternative Classification Algorithms and Implementations 79

12 Preprocessing and Context Representation **81**

12.1 Preprocessing 81

12.2 Tree-based Context Representation 84

12.3 Tokenization 86

13 Merging Conflicting and Incomplete XML Markup **89**

13.1 Introduction and Motivation 89

13.2 Types of Errors in XML-like Input 90

13.3 Configurable Settings and Heuristics for Repair 91

13.4 Algorithm Description 93

13.5 Limitations 96

13.6 Application in Our Approach 97

13.7 Related Work 98

14 Weakly Hierarchical Extraction **99**

14.1 Introduction 99
14.2 Inheritance Hierarchies of Attributes 100
14.3 Strictly Hierarchical Approach and Related Problems 100
14.4 Weakly Hierarchical Approach 101
14.5 Integration into Information Extraction Approach 102

IV Evaluation 103

15 Evaluation Goals and Metrics 105
15.1 Goals and Limitations of Quantitative Evaluation 105
15.2 Evaluation Methodology 106
15.3 Evaluation Metrics 107

16 Text Classification Experiments 109
16.1 Introduction 109
16.2 Text Classification Setup for Spam Filtering 110
16.3 Experimental Results on the SpamAssassin Corpus 110
16.4 TREC Spam Filtering Challenge 114
16.5 Concluding Remarks 117

17 Extraction of Attribute Values 119
17.1 Test Corpora 119
17.2 Evaluation Results for the Seminar Announcements Corpus 120
17.3 Evaluation Results for the Corporate Acquisitions Corpus 124

18 Ablation Study and Utility of Incremental Training 131
18.1 Ablation Study 131
18.2 Utility of Interactive Incremental Training 135

19 Comparison of Tagging Strategies 139
19.1 Idea and Setup 139
19.2 Comparison Results 139
19.3 Analysis 141

20 Weakly Hierarchical Extraction 143
20.1 Experimental Setup 143
20.2 Experimental Results 144
20.3 Concluding Remarks 148

21 Mistake Analysis 149
21.1 Mistake Types 149
21.2 Distribution of Mistakes 150
21.3 Type Confusion 153
21.4 Additional Manual Analysis 155
21.5 Length Analysis 160

V Conclusions **167**

22 Conclusion and Outlook **169**

22.1 Discussion of Results 169

22.2 Summary of Contributions 170

22.3 Future Work 172

Bibliography **175**

A Schema for Augmented Text **185**

B Curriculum Vitae **189**

C Zusammenfassung in deutscher Sprache **191**

List of Tables

2.1 Applications of IR, IE, and Text Mining 16
2.2 Overview of the Selected Approaches and Systems 17

9.1 Example of External Answer Keys 65

10.1 Properties of Tagging Strategies 72
10.2 Labeling Example 72

11.1 SBPH$_5$ Feature Combinations Containing f_5 77
11.2 Features Generated by SBPH and OSB 78

12.1 Regular Expressions Used for Tokenization 87

16.1 Promotion and Demotion Factors 111
16.2 Threshold Thickness 112
16.3 Comparison of SBPH and OSB with Different Feature Storage Sizes . 112
16.4 Utility of Single Tokens (Unigrams) 112
16.5 Sliding Window Size 113
16.6 Preprocessing 113
16.7 Comparison with Naive Bayes and CRM114 114

17.1 Results on the Seminar Corpus 121
17.2 System Comparison on the Seminar Corpus (F-measure) 123
17.3 Results on the Acquisitions Corpus 125
17.4 System Comparison on the Acquisitions Corpus (F-measure) 128

18.1 Ablation Study: Seminar Announcements 132
18.2 Ablation Study: Corporate Acquisitions 133
18.3 Results with Incremental Feedback 136
18.4 Incremental Feedback: User Effort for Correcting the “Training Set” . 137
18.5 Incremental Feedback: User Effort for Correcting the “Evaluation Set” 138

19.1 F-measure Percentages for Incremental Training 140
19.2 F-measure Percentages for Batch Training 140
19.3 Incremental Training: Significance of Changes Compared to *IOB2* . . 141
19.4 Batch Training: Significance of Changes Compared to *IOB2* 142

20.1 Recall Reached by Supertype Recognizers on Subtype Answer Keys . 144

21.1 Seminar Corpus: Length Distribution of Answer Keys 162
21.2 Acquisitions Corpus: Length Distribution of Answer Keys 165

List of Figures

3.1 Tasks to Be Handled 21
3.2 Architecture of a Typical IE System 22
3.3 Sample Interface: Information Extraction from E-Mail Messages . . . 26

9.1 Sample Input Text 63
9.2 Sample Text with Inline Annotations 64

12.1 Partial DOM Tree of a Simple HTML Document with Linguistic Annotations 82
12.2 Processed File from the Seminar Announcements Corpus 83
12.3 Inverted Subtree of the Elements Considered for a Context Representation 84

16.1 Learning Curve for the best setting (Winnow$_{1.23,0.83,5\%}$, 600,000 features, OSB$_5$) 115
16.2 ROC curve for the best filters (Source: [Cor05, Fig. 2]) 117

17.1 Results on the Seminar Corpus 122
17.2 Seminar Corpus: Precision and Recall Improvements 123
17.3 System Comparison: F-measure Averages on the Seminar Corpus . . . 124
17.4 Results on the Acquisitions Corpus 126
17.5 Acquisitions Corpus: Precision and Recall Improvements 127
17.6 System Comparison: F-measure Averages on the Acquisitions Corpus . 128

18.1 Ablation Study: Seminar Announcements 133
18.2 Ablation Study: Corporate Acquisitions 134
18.3 Incremental Feedback: Learning Curve (average precision, recall, and F-measure on all documents processed so far) 137
18.4 Incremental Feedback: Correct, Missing, and Spurious Predictions in the "Training Set" 137
18.5 Incremental Feedback: Correct, Missing, and Spurious Predictions in the "Evaluation Set" 138

20.1 Seminar Corpus: Inheritance Hierarchy 143
20.2 Acquisitions Corpus: Inheritance Hierarchy 144
20.3 Seminar Corpus: F-measure Results 145
20.4 Acquisitions Corpus: F-measure Results 146
20.5 Acquisitions Corpus: Collapsing Short and Long Names 147
20.6 Seminar Corpus: Temporal Predictions 147

21.1 Seminar Corpus: Mistakes Combinations 150

21.2 Seminar Corpus: Distribution of Mistake Types 151
21.3 Acquisitions Corpus: Mistakes Combinations 152
21.4 Acquisitions Corpus: Distribution of Mistake Types 153
21.5 Seminar Corpus: Confusion Matrix (expected type→predicted type) . 153
21.6 Acquisitions Corpus: Confusion Matrix (expected type→predicted type) 154
21.7 Seminar Corpus: Precision and Recall by Token Length 161
21.8 Seminar Corpus: F-Measure by Token Length 161
21.9 Seminar Corpus: Weighted Averages by Token Length 162
21.10Acquisitions Corpus: Precision and Recall by Token Length 163
21.11Acquisitions Corpus: F-Measure by Token Length 164
21.12Acquisitions Corpus: Weighted Averages by Token Length 164