1 Introduction

1.1 Motivation and Goals

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 goal 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.

Researchers working in the area of *text understanding,* one of the precursors of information extraction (cf. Sec. 2.1), aimed at creating a complete formal representation of the contents of a text, but this aim has been found over-ambitious and impossible to realize. To avoid this trap, IE requires a predefined output representation (target schema) and only searches for facts that fit this representation. All other information contained in input texts is simply ignored, as are aspects of language that resist formalization, e.g., the intentions and moods of the authors.

The goal of this thesis has been the development and evaluation of a trainable statistical IE system. The approach is based on two assumptions that so far have been largely ignored by other approaches. One assumption is that *"Systems will be used."* Typical trainable IE systems require to be batch-trained from a set of annotated training texts. The resulting statistical model can be used to propose extractions from other (similar) texts, but it cannot be changed without being rebuilt from scratch. In scientific contexts (including this work), the proposed extractions are only used to evaluate the approach, they do not serve any other purposes. However, for many "real-life" applications, automatic extractions will be checked and corrected by a human revisor, as automatically extracted data will always contain errors and gaps that can be detected by human judgment only. This correction process continually provides additional training data, but batch-trainable algorithms are not very suited to integrate new data, since full retraining takes a long time. To address this issue, this approach supports *incremental training* as an alternative to batch training, allowing successive refinement of an existing statistical model by dynamically adapting it to new training data.

The second new assumption is that *"Structure matters."* While typical IE approaches consider a text as a sequence of words, this approach represents texts by tree structures. This allows considering both implicit linguistic structure and explicit markup information in a unified way. Actually, this is more a conjecture than an assumption, since the effects of using such additional structural information in our system will be evaluated.

Many IE systems are quite monolithic. This makes it hard to find out what causes performance differences between systems since it is impossible or impractical to ex-

change parts of a system. The system developed for this thesis is meant to be usable as a flexibly adjustable framework. All components of the system can be adapted or replaced for purposes of comparing alternative approaches or for introducing improvements without being forced to start from scratch. Some such comparisons have already been conducted as part of this work, for other components this remains as future work.

More specific aims and requirements to be fulfilled by this work will be presented in Chapter 7 after the area of information extraction has been covered in more details. An example motivating the use of information extraction and the advantages of an incremental setup in more detail will be given in Chapter 3.

1.2 Contributions

The core contributions of my thesis[1] lie in four areas:

- I have introduced new functionality not supported by current (statistical) IE systems, especially by designing and implementing an IE system that is suitable for *incremental training* and thus allows a more interactive workflow (following the *"Systems will be used"* assumption given above; cf. Sec. 3.4, Chap. 11, and Sec. 18.2).
- I have designed a generic framework for statistical information extraction that allows modifying and exchanging all core components (such as classifier, context representations, tagging strategies) independently of each other (cf. esp. Chapters 10–12 and 17). I have performed a systematic analysis of switching one such component, namely the *tagging strategies*, describing and evaluating the various tagging strategies that can be found in the literature and also introducing a new one (cf. Sec. 10.2 and Chap. 19).
- I have explored several new sources of information as a way of improving extraction quality. Especially I have introduced rich tree-based context representations that can utilize document structure and generic XML markup (following the *"Structure matters"* conjecture) in addition to the more conventional linguistic and semantic sources of information (cf. Chap. 12 and Sec. 18.1). I have also investigated approaches of integrating hierarchical structures of data such as inheritance hierarchies into statistical IE systems (cf. Chapters 14 and 20).
- I have performed a detailed evaluation of the resulting system on two of the most frequently used standard IE corpora that cover a broad range of the challenges that an IE system may encountered. The evaluation has included an ablation study measuring the influence of various factors on the overall results. It has also included an analysis of the utility of incremental training for reducing the human training effort and an analysis of the kinds of mistakes made by my system and their likely causes (cf. Part IV).

[1] Throughout this chapter, I use the personal "I"-form, while in the rest of this work the conventional "we"-form will be used.

In addition to these core contributions, I have realized several side contributions that resulted from accomplishing the main goals of the thesis, even though they have not been the primary focus of this work:

- Together with my colleague Peter Siniakov, I have established a comprehensive overview of the current state-of-the-art in information extraction, describing and analyzing relevant approaches and providing a classification of types of adaptive IE systems (cf. Part I and [Sie05b]).
- To prepare input documents for creating the tree-based context representations mentioned above, it is necessary to combine different kinds of possible overlapping markup in a single DOM tree structure. For this purpose, I have developed a merging algorithm that can repair nesting errors and related problems in XML-like input (cf. Chap. 13).
- As the core of the classification-based IE approach, I have implemented a generic classifier that turned out to be extremely suitable for other tasks such as spam filtering too. 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 renowned *Text REtrieval Conference (TREC)* (cf. Chapters 11 and 16).

1.3 Outline of this Work

Part I introduces the field of IE and discusses related work, presenting and analyzing the main types of approaches to IE.

Part II analyzes the requirements and desiderata an approach to IE should fulfill. It highlights the assumption underlying the development of the chosen approach and the field of IE in general. The final chapter of the part describes the target schemas and the kinds of input texts the algorithm should be able to handle as well as the desired output format.

Part III describes the architecture and the components of an IE system that fulfills the requirements developed in Part II, including an extension for weakly hierarchical extraction that allows utilizing more generic attributes extracted previously.

Part IV contains a detailed evaluation of both the standard system and extended or modified variants.

The concluding Part V discusses the reached results and expounds open issues and possible future work.

1.4 Acknowledgments

First of all I want to thank my first supervisor, Prof. Heinz F. Schweppe, for his invaluable support and guidance during the preparation of my thesis. Obviously, this thesis would not have been written without him. I am also deeply grateful to my second supervisor, Prof. Bernhard Thalheim, for his helpful comments and suggestions.

I have been lucky to be a member of the Database and Information Systems Group at the Freie Universität Berlin, and I am grateful to everybody in the group for the

good atmosphere. Especially I would like to thank the members of the former *FEx Project*, Peter Siniakov and Heiko Kahmann, for many inspiring discussions and lots of fun.

The financial support of the Berlin-Brandenburg Graduate School in Distributed Information Systems made this thesis possible. Moreover, the graduate school was an important forum for evaluating and improving my ideas. I would like to thank the other graduates for feedback and fun, and the professors for their healthy criticism and occasional clemency.

Finally, I want to thank the other members of the informal text classification and spam filtering team, William S. Yerazunis, Fidelis Assis, and Shalendra Chhabra. This spontaneously formed international team proved to be more stable and persistent than many official projects and has provided a prime example of international cooperation across three continents (or four, if you consider countries of origin).