Part I

Information Extraction: Problems and Perspectives

1 Introduction

Language is the source of misunderstandings.

Antoine de Saint-Exupery

There are things unique to humans that astonish and fascinate at the same time. One of them is the human language, admirable for its richness, complexity and ability to adapt to different cultural and social environments. But as valuable from cultural and aesthetic point of view human language is as challenging it is to grasp it scientifically, to formalize and make it manifest for computer processing.

The area of natural language processing (NLP) gained a lot of attention at the end of 1970's significantly advancing in many research directions such as speech recognition, text understanding, building grammars and conceptual models for natural language. Too optimistic expectations resulting from fast success were not fulfilled and the main goal – understanding and communication in natural language – still remains out of the scope of modern research. In the text-based NLP the focus is being more and more relocated towards solving less complex problems to accomplish very useful practical tasks in text processing and analysis. One of the most promising efforts in this area is *Information Extraction* (IE).

1.1 Information Extraction

Information extraction builds the bridge between the evolutionary aspects of language development and the algorithmic approach to the language. IE is one of the most advancing efforts to exploit computational capabilities, accurateness and correctness of machines for accomplishing elaborate, often tedious task of searching for, analyzing and identifying desired information.

From time immemorial people are used to express and communicate their knowledge, thoughts, emotions by natural language. If the information is not just of momentary interest and should reach a wide audience it is usually communicated as written language. The modern technical innovations did not notably change the human behavior. The reason is probably that it is more comfortable for people to both formulate and receive the information in their natural language rather than inventing new or using alternative communication forms.

Internet, mass media, scientific literature contain and continuously produce huge amount of information that is hidden in natural language texts and stored in digital form. This information can hardly be immediately accessed and processed by computers while human access requires 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 and storing it in appropriate form for future use. The notion of text mining is sometimes used interchangeably with the notion of IE. The goals of information extraction, however, are typically more specific so that we define it as the transformation of facts expressed in natural language to a given, formal, properly defined target structure. The difference to the broader notion of information extraction task should therefore be underlined where the accent is made mainly on the text processing stage and the target representation is less relevant. Information extraction can therefore be regarded as a subset of text mining focusing on more rigidly structured representation forms.

1.2 Related Research Areas

Trying to solve simpler but in practice very relevant problems than text understanding and setting more ambitious objectives than retrieval of relevant documents information extraction can be settled between the scientific fields of *text understanding* and *information retrieval*. The first IE systems were initiated by the DARPA and US Navy in the 1980's [Tur06] as it has become obvious that the research in text understanding would not fulfil the high expectations in the next time. Text understanding has the goal to obtain and represent the complete knowledge comprised by the text in order to be able to answer any question related to the text. IE restricts the range of questions to a small relevant subset.

Main difference between IE and information retrieval is that the result set is not documents or parts of documents containing the relevant information, but the text pieces that immediately express this information and are semantically typed by the attribute whose value they represent. The text models underlying IR and IE differ significantly too. While IR model considers texts as unstructured word sequences [BY99], the IE model takes advantage of layout, syntactic structures and semantic information. However, IR and IE can be combined in various ways. IR can be used to identify relevant documents where an IE system can look for information on a finer level. On the other hand, results of IE can be used for information retrieval replacing for example standard keyword queries by SQL queries on a database with extracted facts. Hence IE may be useful as a preparatory step for information retrieval as well as a refinement of retrieval. Evaluation of the quality of both IR and IE is not straightforward. The standard correctness criteria cannot be applied to the results in both cases: correctness of a result set cannot be judged by a binary decision. IE borrows the quantitative evaluation metrics *recall* and *precision* from IR to assess the goodness of extracted items.

As we argued above, text mining [Wei05] is closely related to IE, because similar goals are pursued. The essence of text mining is to discover recurring and predictive text patterns in order to derive new information by applying data mining techniques to natural language texts. IE methods can be employed in text mining to extract facts from the unstructured text to a database or other structured representation. In a second step usual data mining techniques can be applied to

the resulting database to discover new relationships in the data [Nah00].

1.3 The Problem of IE

However, the problem this research is devoted to is not immanent in information extraction. Taking a more abstract view, the essence of the problem is the search and the acquisition of information from natural language texts after the desired information has been a priori specified. Information extraction provides just the frame, that is defines additional formal restrictions for the task stating the requirements for the specification of information and further processing of acquired information. The problem manifests itself in many other tasks: e.g. when answering an in advance defined query or question or just localizing the desired information in the text. The following questions are helpful to better comprehend and approach the problem and are partially discussed in the following sections:

- ▷ How is the information to be extracted defined?
- ▷ What are the elementary units of information?
- ▷ What advantages, positive restrictions does the a priori specification of information imply?
- ▷ How does the information manifest itself in the natural language, what components of the language bear the information?
- ▷ What linguistic properties help to localize the desired information?
- ▷ Is it necessary to understand the text in order to identify information?

Human beings solve this problem by understanding the text and deducing the desired information from the semantic model they build by logically connecting different text parts with each other and their background knowledge. Complete understanding of a text requires a great deal of background and world knowledge. Creation of adequate knowledge bases is very tedious and often not feasible. One of our central assumptions is that it is not necessary to completely understand the text to acquire a priori specified information. It is justified by the fact that the goals of information extraction are less ambitious than those of text understanding. An information extraction system does not have to be able to answer any questions after having the text processed, but only those questions specified formally before processing the text. To a much larger degree than text understanding IE can take advantage of statistical properties of the language learning frequently recurring patterns, relevant syntactic properties etc.

The solution of IE problem can significantly benefit from the restriction to certain items of interest – the *target structure* – defined in advance that allows the utilization of supervised learning algorithms. The simplest abstract specification of desired information is a template that consists of a number of slots that should be filled with elementary extractions we sometimes call *facts*. Reduced to slot filling the task of IE consists in identifying all relevant text fragments and assigning them to an appropriate slot. However, in this case the relation of single slots to each other is neglected and a complex entity consisting of several slots cannot be extracted if there are at least two instances of this entity in a document. If, for example, during extraction of personal data two first and two last names are encountered in a text, the correct assignment of text fragments to

the corresponding name part will not be sufficient to extract a person since first and last names are not related to each other.

Using database relations as target structure involves that the relevant units of information cannot be extracted independently from each other and have to be combined in a relation tuple. In this case facts have the form $P(a_1, ..., a_n)$ where P is a predicate or relation and $a_1, ..., a_n$ are typed values. We use the notion of *attribute* as the value type according to the terminology of databases. Beside the identification of relevant content an important aspect of IE is concretized here: the consolidation of belonging together elementary information items (facts) distributed in the text. Determining how different attribute values are related to each other is a challenging problem that extends the classical notion of IE as slot filling to recognition of complex facts and entities.

Hence, the extracted text parts cannot be regarded independently from the target structure bringing up the question what the extracted information eventually is. Text fragments that are identified as relevant and assigned to attributes of the target structure do not necessarily have to correspond with the attribute values stored in the database. The text fragments may be normalized according to the expected format (e.g. representation of dates and times) or some predefined (e.g. enumerated) values may be derived from the extractions. Some identified facts may appear in text more than once or already exist in the database. In this case different instances could be merged (instance unification) before being inserted in the database. The actual content of the database may therefore be considered as the extracted information.

The problem of IE can be defined differently depending on the input requirements, notion of extracted information and the spectrum of tasks IE includes. In our research we concentrate on the classical tasks of IE – identification and extraction of attribute values – presenting an algorithm that is capable of relation extraction without actually evaluating it. A predefined target structure in form of a template or a database schema is presupposed. The text fragments that were identified as relevant are regarded as extracted facts. They constitute the attribute values and are not normalized or altered.

1.4 Why is the problem of IE not solved?

Since more than a decade IE has been being an evolving research field giving rise to exploration of many new methods and numerous systems. However, IE is still far away from a mature technology that can be easily employed in commercial systems. Satisfactory extraction accuracy obtained on not very challenging corpora used in research cannot be achieved in many real world domains. Main difficulties of finding and extracting information are diversity and ambiguity of natural language. Natural language offers plenty of possibilities to express a certain fact. As much as this diversity contributes to the language richness as difficult it is to handle by automatic language processing, since the processing system has to be able to recognize numerous possibilities of fact expression. Besides, a certain text fragment can have several semantics and syntactic roles. Because of morphological, lexical, structural and syntactic ambiguities hidden in natural language it is very difficult to determine the correct interpretation of sentences or even words. Since the linguistic processing precedes actual fact extraction, mistakes such as incorrectly parsed sentence or incorrectly determined part of speech made on this stage have a considerable effect on the extraction

results. Semantic ambiguities include the possibility of assignment of the same text fragment to different attributes, expression of several facts in one phrase, uncertainty in the clauses when it has to be determined whether the information contained in reported speech or expression of assumptions etc. is factual, and many more.

However, not every natural language text is equally ambiguous or variable in forms of expression. The spectrum of natural language documents reaches from highly structured form-like texts that are not fully grammatical and sometimes telegraphic in style to philosophical texts with many intratextual relations. An important prerequisite for the practical use of current IE algorithms is homogeneous, stylized texts, which confine the diversity and ambiguity of natural language to a level that can be handled by today's machine learning techniques. Our approach is primarily targeted at information extraction from texts written in technical language. It is based on the assumption that technical languages are distinguished by restricted linguistic forms of expression. The regularities and conventions that are inherent in a certain technical language can be learned by an IE system to identify relevant content. An interesting aspect of the research we also address is the effectiveness of the approach in domains where more challenging, for example journalistic, language style is used.

While the complexity of natural language makes it difficult to reliably identify relevant text fragments, recognition of relations is complicated by additional factors. One of the major difficulties is the object identification. In the relational model relation tuples are identified by a primary key that consists of relation attributes. If the primary key contains attributes that have to be extracted, the extraction may sometimes be inhibited because a value of a primary key attribute is not found by the system even though all other attribute values are determined correctly. Artificial keys (i.e. a numeric attribute that is not extracted) help to solve this problem but heighten the risk of incorrect extraction of attribute values. Since the existence of a real identifier of a relation tuple is not presupposed any more, any, also wrongly extracted single attribute value can be stored as a relation tuple with some artificial key. Artificial keys lead therefore to the dispersal of the extracted data to many partial relations. Besides, the expedience of partial relation extractions enabled by artificial keys, even if they are correct, is quite questionable in many domains.

If the attribute values are widely scattered in the text, IE systems often fail to assemble them in a correct relation because no direct context or any other plausible evidence of their relation except for deep semantic connection is available. Because of these complicating factors the majority of approaches to IE do not perform relation extraction leaving it as future work. Current research in relation extraction usually circumvents the problems described above focusing only on mere detection of binary relations and neglecting the actual transfer of text fragments to the database.

1.5 Outline of this Work

Among numerous solution proposals for the problem of IE rule-based and statistical approaches constitute two predominant classes. While the latter adapt well known statistical learning techniques to the IE problem, the former try to learn explicit rules how to find and extract desired information. Basing our approach on the assumption that there are certain patterns for the expression

of certain information and favoring the declarative representation of acquired linguistic knowledge we develop a new rule-based approach to IE significantly extending the common rule-based model and introducing a novel learning algorithm for extraction rules. Below we summarize some of the core aspects of our work.

The majority of rule-based approaches employs very simple linguistic patterns (often restricting the pattern model to the subset of regular expressions). In this work we examine the utility of a comprehensive, context-free, fully recursive, sentence-based pattern model that is capable of integration of additional structural and linguistic information and XML markup. The tradeoff here lies between the gain of expressivity and the growing complexity of the learning algorithm.

Our endeavor to embed many sources of different (e.g. linguistic and structural) information involves that the free texts cannot be used as the basis for matching linguistic patterns. Instead patterns have to be matched with XML documents that contain all obtained information about texts as XML elements. Since the extraction patterns act as XML queries, the question is worth consideration whether IE task can be facilitated by XML query processing. We demonstrate the equivalence of our pattern specification language and the corresponding XML query language modeling the extraction task as pattern unification on XML documents.

The big advantage of adaptive approaches to IE is their ability to learn the extraction rules from a small set of annotated training texts. The crucial question is how precise the characteristic features of relevant information can be captured by the learning algorithm. In our work we examine the effectiveness of inductive learning in combination with our pattern model. We introduce three generalizing heuristics that are responsible for deriving reliable and general extraction rules and are formally defined on the proposed pattern language.

As opposed to simple clustering rule generalization by merging similar rules is supposed to improve their quality. To reliably identify common features a rule similarity metric has to consider the similarity of extractions, their co-occurrence and their context. Since extraction rules consist of hierarchical structures, a serious problem is whether such a similarity metric can be efficiently computed.

During the induction of extraction rules the component that assesses the goodness of induced rules (rule validation) plays a major role for the performance on the test texts. One of the most difficult challenges is to keep the optimal balance between precision and recall. A promising and yet unattempted solution is to identify parameters of extraction rules that determine precision and recall and optimize the set of induced rules for maximum F-measure. The optimized parameters can be used not only at the training but also at the extraction stage providing a confidence measure for the validity of an extraction.

As mentioned above, one of the biggest problems for IE is the diversity of natural language. Can this diversity at least partially be handled by generating a thesaurus for regarded application domain? We investigate this question developing an algorithm for synonymy recognition and using it for generalization of lexical expressions in linguistic patterns.