9 Case Studies

9.1 Introduction

In the foregoing chapters we focused on the history of information extraction, on the current extraction technologies and their evaluation. In this chapter, it is time to illustrate these technologies with real and recent case studies, to summarize the capabilities and the performance of these systems, and to draw the attention to the bottlenecks that need further research. Furthermore, we will sum up the tasks in which the extraction technology is integrated and specifically focus on information that is relevant in a retrieval setting.

Information extraction technology is integrated in a variety of application domains and many different tasks are being implemented.

Information extraction from news texts has considerably been studied in the past research. Information on worldwide events such as natural disasters, political events or on famous persons is commonly identified in the documents.

Another application domain where information extraction technology is in full expansion is the *biomedical domain*. In this domain, extraction has become a necessary technology in order to master the huge amounts of information, much of which is in the form of natural language texts.

A third domain, which currently gives a strong impetus to the development of information extraction technology, is *intelligence gathering*. After the September 11 attacks, police and intelligence services are eager to find and link information in the bulks of private e-mails, phone conversations and police reports, and in public sources such as news, chat rooms and Web texts.

In the *economic and business domain*, there is a large interest in extracting product and consumer information from texts found on the World Wide Web, to monitor mergers and transactions, and to identify consumer sentiments and attitudes. These texts usually carry some structure marked with HTML (HyperText Markup Language) or XML (Extensible Markup

Language). In the business domain, one is also interested in extracting information from technical documentation.

In the *legal domain* we see a large demand for information extraction technologies, especially for metadata generation and concept assignment to texts, which could be used for case-based reasoning. Notwithstanding this need and the huge amounts of written texts that are available in legal databases, information extraction is not very much researched in the legal domain. Moreover, the results of the rare studies in information extraction leave room for a lot of improvements.

Finally, information extraction *from speech and informal sources* such as e-mail and spam poses additional difficulties that are the focus of current research.

The performance measures that accompany our case descriptions are only indicative because the evaluation settings (corpora, semantic classes, selected features) usually differ. The aim is to give the reader an estimate of the state of the art performance. We refer to the literature for details on the evaluations. Unless stated otherwise, the results regard the processing of English texts.

The above list of extraction tasks is far from exhausted and is only inspired by information extraction from text. Any information medium that is consulted by humans is or will be eventually accessed with information extraction technologies.

Before discussing the different application domains of information extraction, we will give some general remarks on the generic versus domain specific character of the extraction technology.

9.2 Generic versus Domain Specific Character

In the previous chapters we have described the technologies on a very general level and treated fairly *generic extraction tasks* such as named entity recognition, noun phrase coreference resolution, semantic role recognition, entity relation recognition, and timex recognition and resolution. These chapters show that the information extraction algorithms and methods can be transposed to many different application domains. However, within a certain domain the extraction tasks become more refined (e.g., domain specific entities are extracted) as each domain carries additional *domain specific semantic information*. The domains also handle *specific text types or genres* (e.g., newswires, news headlines, article abstracts, full articles, scientific reports, technical notes, internal communiqués, law texts, court decisions, political analyses and transcriptions of telephone conversations).

Variations between subject domains mainly come down to the use of a specialized vocabulary and of certain domain specific idiomatic expressions and grammatical constructions, besides the vocabulary, expressions and constructions of ordinary language. For instance, biomedical texts use many domain specific names, while legal texts are famous for their use of lengthy, almost unreadable grammatical constructions.

Variations between text types mainly regard the rhetorical and global textual features. The former includes the use of specific rhetorical markers, of specific forms of argumentation or causality, of the directness of the message, and the underlying goal of the text. The latter includes parameters such as text length, use of typography and specific rules as to text formatting. For example, a news feed will almost always be a short text that wants to inform the reader in a neutral and direct tone that a certain noteworthy event took place somewhere in the world. It will contain a headline (which usually summarizes the event described in the text body and is often capitalized or typographically distinct from the rest) and a small number of very short paragraphs. Scientific journal articles are usually longer; they do not necessarily describe a noteworthy event, but rather the result of scientific research; they do not simply want to convey something, but try to convince the reader that the research described in the article is scientifically relevant and sound; and they do that – or at least are supposed to do that – by using some form of rational argumentation. In their most simple form, they are organized into a number of subsections, each of which has a subtitle and is subdivided in a number of paragraphs. The articles are preceded by a title that is indicative of the content and an abstract containing a short overview of the article, and consist of a main body that has a topic-argument-conclusion structure.

As a result the domain specific extraction tasks rely on domain specific and text type specific contextual features, often demanding different levels of linguistic processing – sometimes domain adapted linguistic processing – in order to compute the features values. In addition, an *ontology of domain specific semantic labels* accompanies the information phenomena.

Although it is not easy to choose and model the right features and labels for the extraction task, the underlying *technology and algorithms* – especially the pattern recognition techniques – for information extraction are fairly *generic*. The difficulty in defining good features is one of the reasons why information extraction has been popular in a restricted semantic domain operating on a certain text type. Nowadays we have at our disposal many natural language processing tools that can be applied for feature selection. A completely domain independent information extraction system does not exist because of the reliance on a rich variety of features, but recent trends in information extraction increasingly stress the importance of

making extraction components as generic as possible, especially the underlying algorithms and methods.

These findings make information extraction also interesting for information retrieval from both specific document collections and collections that cover heterogeneous domains and text types, such as found on the World Wide Web.

9.3 Information Extraction from News Texts

Information extraction from news is well developed through the *Message Understanding Conferences* (*MUC*) of the late 1980s and 1990s, sponsored by the US Defense Advanced Research Projects Agency (DARPA). Many of the MUC competitions involved the extraction of information from newswires and newspaper articles. For instance, MUC-3 and MUC-4 involved the analysis of news stories on terrorist attacks. MUC-5 included texts on joint ventures, while MUC-7 identified information in news on airplane crashes and satellite launch events. Each of the MUC conferences operated in a specific domain, though the MUC experiences laid the foundations for many generic information extraction tasks (e.g., noun phrase coreference resolution, named entity recognition) and they showed that the technology developed could be easily ported to different domains. The MUC competition focused also on finding relations between different entities that form the constituents of an event and that fill a template frame, e.g., time, location, instrument and actors in a terrorist attack. Typically in news actors and their relations (who did what to whom) and the circumstances (e.g., location, date) are identified.

Currently, the *Automatic Content Extraction* initiative (*ACE*) of the National Institute of Standards and Technology (NIST) in the US develops content extraction technology to support automatic processing of human language in text form. One of the source types involves newswire texts. An important goal of this program is the recognition of entities, semantic relations and events. The entities include persons, organizations, geographical-political entities (i.e., politically defined geographical regions), localization (restricted to geographical entities), and facility entities (i.e., human made artifacts in a certain domain). In addition, relations between entities are detected. They include within and across document noun phrase coreference resolution, cross-document event tracking and predicate-argument recognition in clauses. In the frame of the above competitions valuable annotated test corpora were developed.

Named entity recognition – and more specifically recognition of persons, organizations and locations – in news texts is fairly well developed, yielding performance in terms of F-measure[1] (see Eq. (8.5)) above 95% for different machine learning algorithms (e.g., maximum entropy model, hidden Markov model) (e.g., Bikel et al., 1999). The performance of named entity taggers on written documents such as Wall Street Journal articles is comparable to human performance, the latter being estimated in the 94-96% F-measure range. This means that relevant features are very well understood and that the patterns are quite unambiguous.

The best results of *noun phrase coreference resolution* are obtained with decision tree algorithms (F-measure of 66.3% and 61.2% on the MUC-6 and MUC-7 data sets, respectively) more specifically for the decision tree algorithm (C.4.5) (Ng and Cardie, 2002) and F-measures in the lower 60% when resolving coreferents with weakly supervised methods (Ng and Cardie, 2003). F-measures are here computed based on the Vilain evalution metric for recall and precision (see Eqs. (8.12) and (8.13)). The results show that noun phrase coreference resolution in news texts is far from a solved problem.

Cross-document noun phrase co-reference resolution research investigates both the detection of synonymous (alias) names (mentions with a different writing that refer to the same entity) and the disambiguation of polysemous names (mentions with the same writing that refer to different entities). Li et al. (2004) report results of 73% in terms of F-measure, when applying a generative language model on 300 documents from the New York Times for the former task (cross-document alias detection of people, location and organization mentions while ignoring anaphoric references). For the disambiguation task, the same authors obtain an F-measure close to 91% under the same settings. Gooi and Allan (2004) report best results in terms of F-measure (obtained with the B-CUBED scoring algorithm for recall and precision, see Eqs. (8.16) and (8.17)) of more than 88% by clustering terms in contextual windows on the John Smith corpus with the aim of disambiguating the name John Smith across different documents.

With the ACE corpus of news articles (composed of 800 annotated text documents gathered from various newspapers and broadcasts), Culotta and Sorensen (2004) obtain promising results on an *entity relation recognition* task by using different dependency tree kernels. The kernel function is used as a similarity metric. Given a set of labeled instances, the method determines the label of a novel instance by comparing it to the labeled instances

[1] Unless stated otherwise, F-measures (see Eq. (8.5)) refer here to the harmonic mean, where recall and precision are equally weighted (also referred to as F_1-measure).

using this kernel function. For a binary classification with a support vector machine (SVM) the tree kernel and the combination of the tree (contiguous or sparse kernel) and bag-of-word kernel outperform the bag-of-word kernel by F-measures between 61% and 63% versus 52%. Precision is quite good (in the lower 80%), but is tempered by the rather low recall values (ca. 50%). The 24 types of relations used (e.g., semantic roles that indicate e.g., part-of relation, specific family and affiliation relations) have a quite different distribution in the data set. 17 of the relations have less than 50 annotated instances, being an important cause of the low performance in terms of recall.

The lack of patterns in the training data is an important, but not the sole cause of a low recall value. Another problem is implicit information, i.e., there is information that is not made explicit in the stories, but is understood by human readers from its context.

In addition, news stories are often of the narrative genre. They are very well suited to establish the timeline of the different steps in an event or the timeline of different events. Research in the *recognition and resolution of timexes* is only in its infancy, but becomes an important research topic. Recognition and resolution of timexes have to deal with ambiguous signaling cues and content left implicit in the text (e.g., the time order of certain content is not explicitly expressed in a text or is lacking across texts, but the logical order is easy to infer based on world knowledge).

Not all the news stories are of narrative nature. Many of them are also opinion pieces or interweave opinions into the events. Information extraction technology used in *opinion extraction* of news is limited (Grefenstette et al., 2004), although there is an extensive literature on sentiment or attitude tracking (see below).

Information extraction from news is important in question answering retrieval. Information extraction from the text of news is increasingly used to annotate accompanying images and video (Dowman et al., 2005), and there is no doubt that these annotations will play a valuable role in the retrieval of news media. Current research focuses on aligning content recognized in news media across text and images in order to obtain well-documented answers and summaries to information questions.

9.4 Information Extraction from Biomedical Texts

Among the application domains of information extraction, the biomedical domain is currently the most important. This is due to the large amount of

biological and medical literature that exponentially grows every day and the necessity to efficiently access this information.

A first source of information regards *patient reports*. There have been efforts to extract information and consequently encode the information in order to use it in data mining, decision support systems, patient management systems, quality monitoring systems, and clinical research.

A second source of information is made of the huge repositories with *scientific literature* in the biomedical domain. Medline alone contains over 15 million abstracts and is a critical source of information, with a rate of ca. 60.000 new abstracts appearing each month.

Different ontologies or *classification schemes* and annotated databases are available, e.g., the functional annotations listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004) and the Gene Ontology (GO) (Ashburner et al., 2000) annotation databases. The Gene Ontology is a large controlled vocabulary covering molecular functions, biological processes and cellular components. An important annotated dataset is the GENIA dataset. Currently the GENIA corpus is the largest annotated text resource in the biomedical domain available to the public. In its current version, it contains 2000 MEDLINE abstracts that are part-of-speech tagged with Penn Treebank POS tags and annotated with biomedical concepts defined in the ontology.

Especially, *named entity recognition* is a very common task because of the absolute necessity to recognize names of genes, proteins, gene products, organisms, drugs, chemical compounds, diseases, symptoms, etc. The named entity recognition is a first step for more advanced extraction tasks such as the detection of protein-protein interaction, gene regulation events, subcellular location of proteins and pathway discovery. In other words the biological entities and their relationships convey knowledge that is embedded in the large textual document bases that are electronically available.

Named entity recognition poses specific problems because of the complex nature of the detection of the boundaries of the entities, their classification, mapping (tracing) and disambiguation. These problems also occur in other application domains, but are usually less pronounced in these domains.

Boundary detection of the named entity is not always easy and its recognition is often treated as a separate classification task. One cannot rely on a simple short type that defines capitalization or other character patterns because of the variety of patterns that refer to the same named entity. Biomedical named entities often have pre-modifiers or post-modifiers that are (e.g., **91 kDA protein**) or are not (e.g., activated **B cell** lines) part of the entity. The names are often solely mentioned or referred to as acronyms (e.g., **NR** = nerve root). Entities are of varying length (e.g., **47 kDa sterol**

regulatory element binding factor, RA). Two or more biomedical named entities can share one head noun by using a conjunction construction (e.g., **91 and 84 kDa proteins**). Biomedical entities are often embedded in one another (e.g., <PROTEIN> <DNA> **kappa 3** </DNA> **binding factor** </PROTEIN>).

Commonly used features in the classification task are orthographic features (e.g., use of capitals, digits), morphological prefixes and suffixes (e.g., **~cin, ~mide**), part-of-speech features, head noun words, contextual trigger words such as verb triggers (e.g., **activate, regulate**), head noun words (e.g., **treatment, virus**).

Biomedical names have many aliases (synonym terms, acronyms, morphological and derivational variants), reinforced by the ad hoc use of orthography such as capitals, spaces and punctuation (e.g., **NF-Kappa B**, **NF Kappa B**, **NFkappaB** and **NF kappa B**) and the inconsistent naming conventions (e.g., **IL-2** has many variants such as **IL2**, **Interleukin 2** and **interleukin-2**). On the other hand, names and their acronyms are often polysemous. Although exhibiting the same orthographic appearances, they can be classified in different semantic classes, depending on a given context (e.g., **interleukin-2** is a protein in some context, but can be a DNA in another context; **PA** can stand for **pseudomonas aeruginosa**, **pathology** and **pulmonary artery**). Existing lexico-semantic resources in this domain typically lack contextual information that supports disambiguation of terms. This situation makes that within and cross-document noun phrase coreference resolution is a necessity.

New terms and their corresponding acronyms are invented at a high rate while old ones are withdrawn or become obsolete.

Although the earliest systems rely on handcrafted extraction patterns, current named entity recognition in the biomedical domain use machine learning techniques. The results of a hidden Markov model (Zhang et al., 2004) have an average of 66.5% F-measure for 22 categories assigned of the GENIA ontology. The F-measures range from 80% (category **body-part**) to 0% (e.g., categories **atom, inorganic**). The lack of sufficient training examples in this experiment and resulting low recall are an important factor in the low F-measure for certain categories. Kou et al. (2005) made a comparative study on protein recognition on the GENIA corpus. The results are 66% in terms of F-measure when training with a maximum entropy classifier and 71% when training a conditional random field classifier. The results of the CRF model could be improved by about 1% through an extension of the conditional random fields (SemiCRFs) that enables more effective use of dictionary information as features. Lee et al. (2004) train a Support Vector Machine and consider entity boundary detection and entity classification as two complementary tasks. Tests with the

GENIA corpus yield a best F-measure of 74.8% for the boundary detection task and of 66.7% for the entity classification task. Finkel et al. (2005) use a maximum entropy model for detecting gene and protein names in biomedical abstracts. Their system competed in the Biocreative comparative evaluation and achieved a precision of 83% and recall of 84% (F-measure of 83%) in the open evaluation and a precision of 78% and recall of 85% (F-measure of 83%) in the closed evaluation. In an open evaluation extra resources in the form of gazetteers (i.e., lists of names) or related texts were used.

The detection of entity boundaries in biomedical information extraction is a problem by itself. The identification of the boundaries is difficult because of the diverse use of modifiers such as adjectives, past particles or modifying prepositional phrases, and it is hard to distinguish whether a modifier is included in the named entity or not. Including here statistics on the collocational use of the separate terms as extra features seems useful. Finkel et al. (2005) also stress the importance of correct boundary detection as a way of improving the named entity recognition task in biomedical texts. In their research many errors (37% of false positives and 39% of false negatives) stem from incorrect name boundaries.

Recall (or false negative) errors are caused by patterns not seen in the training set, i.e., the classifier does not know the name and/or contextual pattern. An initial normalization of the training and test examples that is correctly performed seems very useful. Especially a syntactic normalization with syntactic equivalence rules might be helpful. However, this is not always easy. For instance, it is not simple to detect an instance of a coordinated noun phrase where the modifier is attached to only one of the phrases and modifies all of the coordinated members.

Researchers seem to agree that in order to improve named entity recognition in the biomedical domain, we must explore other avenues, including better exploitation of existing features and resources, development of additional features, incorporation of additional external resources, or experimentation with other algorithms and strategies for approaching the task.

The named entity recognition is a first step for more advanced extraction tasks such as the detection of protein-protein interactions, protein-nucleotide interactions, gene regulation events, subcellular location of proteins and pathway discovery. These complex tasks involve relation detection. Current progress in genomics and proteomics projects worldwide has generated an increasing number of new proteins, the biochemical functional characterization of which are continuously being discovered and reported.

Entity relation recognition can be based on hand-built grammars by which the texts are partially parsed. An example hereof is the research of Leroy et al. (2003). They use cascaded finite state automata to structure

relations between individual biomedical entities. In an experiment considering 26 abstracts they obtained 90% precision. Gaizauskas et al. (2000) built an extraction system that heavily relies on handcrafted information resources, which include case-insensitive terminology lexicons (the component terms of various categories), morphological cues (i.e., standard biochemical suffixes) and handcrafted grammar rules for each class. The system is applied for the extraction of information about enzymes and metabolic pathways and the extraction of information about protein structure.

More advanced techniques use machine learning for *protein relation extraction*. Ramani et al. (2005) recovered 6,580 interactions among 3,737 human proteins in Medline abstracts. Their algorithm has three parts. First, human protein names are identified in Medline abstracts using a recognizer based on conditional random fields. Then, interactions are identified by the co-occurrence of protein names across the set of Medline abstracts. Finally, valid interactions are filtered with a Bayesian classifier. A similar approach is taken by Huang et al. (2004) who aligned sentences whose protein names were already identified. Similar patterns found in many sentences could be extracted as protein relations.

Literature-based gene expression analysis is a current research topic. Texts that describe genes and their function are an important source of information in order to discover functionally related genes and genes that are simultaneously expressed. The texts give an additional justification and explanation (Glenisson et al., 2003).

The function of a protein is closely correlated with its subcellular location. With the rapid increase in new protein sequences entering into data banks, textual sources might help us to expedite the *determination of protein subcellular locations.* Stapley et al. (2002) evaluated the recognition of 11 location roles in Medline abstracts and obtained F-measures ranging from 31% to 80% depending on the location class.

Pathway prediction aims at identifying a series of consecutive enzymatic reactions that produce specific products in order to better understand the physiology of an organism, to produce the effect of a drug, understand disease processes and gene function assignment. Complex biomedical extraction tasks aim at predicting these pathways. The information extraction task is similar to detecting an event or scenario that takes place between a number of entities and to identifying how the actions that constitute the scenario are ordered (e.g., in a sequence of reactions of a pathway). This means that the clausal and textual levels of analysis will become relevant and that we will have to resort to event extraction and scenario building technologies to solve this problem. Research on pathway recognition is already done by Friedman et al. (2001).

The overview given here is far from exhaustive. The biomedical literature is full of experiments that report on information extraction from textual sources and on the integration of data extracted from *unstructured texts* with *structured data*. Biomedical information is also increasingly extracted from figures and figure captions.

9.5 Intelligence Gathering

Police and intelligence services are charged with collecting, extracting, summarizing, analyzing and disseminating criminal intelligence data gathered from a variety of sources. In many cases the sources are just plain text. Processing this data and extracting information from them is critical to the strategic assaults on national and international crime. The information is necessary to combat organized criminal groups and terrorists that could threaten state security.

Most criminal data are *structured* and stored in relational databases, in which data are represented as tuples with attributes describing various fields, such as attributes of a suspect, the address of a crime scene, etc. *Unstructured data*, such as free-text narrative reports, are often stored as text objects in databases or as text files. Valuable information in such texts is difficult to access or to efficiently use by crime investigators in further analyses. Recognizing entities, their attributes and relations in the texts is very important in the search for information, for crime pattern recognition and criminal investigation in general. Combined with factual data in databases, the extracted information is very helpful as an analysis tool for the police.

We can make a distinction between open and closed data sources of the intelligence services. The *open sources* are publicly available, have a variable degree in reliability, and include Web pages, files downloadable via the Internet, newsgroup accounts, magazine and news articles, and broadcasted information. *Closed sources* have a secured access and are available only to certain authorized persons. They include police and intelligence reports, internal documentation, investigation reports and "soft" information (i.e., information on suspicious behavior that is noted). The sources are not only composed of texts, but are increasingly of multi-media format. The textual sources are often of multi-lingual nature.

Police forces and intelligence services worldwide start using commercial mining tools, but they are not always adapted to their specific needs. On the other hand, research into the specific demands of extraction systems that operate in this application domain is limited or is not publicly

available. MUC-3 and MUC-4 already covered news articles on the subject of Latin American terrorism. DARPA (Defense Advanced Research Projects Agency) recently started the research program *Evidence Extraction and Link Discovery* (EELD). The purpose of this project is the development of accurate and effective systems for scanning large amounts of heterogeneous, multi-lingual, open-source texts (news, Web pages, e-mails, etc.). The systems should identify entities, their attributes, and their relations in a larger story (scenario) in order to detect crime patterns, gangs and their organizational structure, and suspect activities (e.g., a person **John B** drives a white Alfa Romeo).

In all of the tasks described above, entity recognition is of primordial importance. Entities are first of all the common entities such as person, organization and location names and timexes, but they comprise also car brands, types of weapons, money amounts and narcotic drugs. In addition, it is very important to link the entities to each other, where the link will be semantically typed.

There are very few evaluations of the performance of *named entity recognizers* that operate on police and intelligence texts. Chau and Xu (2005) trained a neural network pattern recognizer combined with a dictionary of popular names and a few handcrafted rules in order to detect and classify the entities of 36 reports that were randomly selected from the Phoenix Police Department database for narcotic related crimes. The reports were relatively noisy: They are all written in uppercase letters and contain a significant amount of typos, spelling errors, and grammatical mistakes. The following entities were recognized: persons (with a precision of 74% and recall of 73%), addresses (with a precision of 60% and recall of 51%), narcotic drugs (with a precision of 85% and recall of 78%) and personal property (with a precision of 47% and recall of 48%). These numbers sharply differ from the precision and recall numbers of the entities extracted from news text. A first reason regards the orthographical and grammatical errors found in these texts. Secondly, entities other than person names, organizations and locations, such as drug names, crime addresses, weapons and vehicles are also relevant to crime intelligence analysis, but they are sometimes more difficult to extract as the contextual patterns are more ambiguous (e.g., **Peter B. gave the Kalashninov to Sherly S. in Amsterdam Centraal.** and **Peter B. gave the Cannabis to Sherly S. in Amsterdam Centraal.).** These additional entity types do not often change names, so that external lexico-semantic resources can easily be used and maintained, unless the entities have code names in the captured messages.

Noun phrase coreference resolution is of absolute importance. Especially, persons and their references need to be tracked in and across docu-

ments. As in any other application domain we have to disambiguate the names and their aliases. Criminals very often use or are referred to by different names, that orthographically might be completely different (e.g., **Peter B.** aliased as **Petro** and **The big sister**) making the task of name tracking a special challenge.

To our knowledge research into relation recognition is very limited. For instance, *extraction of subordination relations* between entities were detected in 1000 intelligence messages in order to construct the hierarchies of organizations (Crystal and Pine, 2005). The relations are detected as explicitly stated connections between two entity mentions (e.g., **Muhammad Bakr al-Hakim is a leader of Iraq's largest Shiite political party** is classified as a leadership relation). The authors report F-scores of 91% for recognition of names, of 83% for entity coreference resolution and 79% for subordination relation detection. Scores assume 50% partial credit assigned to "easily correctable" errors such as misplaced name boundaries (e.g., including a title in a person name). Both in the biomedical and in the police and intelligence domains recognition of relations between entities is important. Whereas in the former domain one could rely on the many instances to affirm the validity of the detected relation, in the police and intelligence domain a single instance of a relationship could be of primordial importance.

Police and intelligence services are also interested in building *a profile of an entity* based on a corpus of documents. The extraction system should collect entity-centric information based on coreference and alias links. Different attributes of an entity should be detected (e.g., Peter B. has **red** hair; the car has a **damaged back light**).

As seen in the previous section, *extraction, resolution and ordering of temporal expressions* (timexes) are valuable tasks when processing news stories. In the police and intelligence domain, they are of primordial importance. Temporal information is crucial to link persons, to link persons to events, to track people and the events they participated in and to link events. Extracting temporal information from the texts involves the recognition of the temporal expressions and the classification of the temporal relations: Anchoring to a time (e.g., when exactly an event takes place is often relative to the time of detection of a crime and is often vaguely described, e.g., **Sometime before the assassin meeting**, the two men must have flown into Chicago), ordering between temporal events, aspectual relations (detecting the different phases of an event) and subordinating relations (events that syntactically subordinate other events).

In police and intelligence settings the *recognition and resolution of spatial information* is also very valuable in order to, for instance, link persons to events. Processing spatial information goes beyond the recognition of location names, but includes also the resolution of spatial references (e.g., up there) and the placing of persons and objects (e.g., car) in a spatial context (e.g., the city of **Brussels** is mentioned in the beginning of the text and the **car jacking** mentioned at the end of the text: does the carjacking take place in Brussels?). Spatial information is often vague, ambiguous and hard to extract.

Extraction of temporal and spatial relations demands correctly annotated corpora. These are not easy to obtain given the ambiguity and vagueness of the language. Some resources for evaluation and training are available. There are the TimeBank data in which timexes and their relations are annotated. In addition, corpora labeled with temporal information gathered for the task of information retrieval become available (e.g., the AQUAINT corpus).

The entities in which police and intelligence services are interested, are often the building stones of an *event description*. Many different types of events are interesting. The "what", "who", "where", "when", "frequency" of a meeting or a crime can be extracted. The "who", "when", "length", "content" and "frequency" of a phone call can be identified. Other types of events are possible (e.g., delivery, travel) of which information has to be collected.

In this domain *scenario* or *script extraction* is relevant in order to classify a set of consecutive actions (e.g., describing a set of actions as a bank robbery) or content that is linked with types of rhetorical relations (e.g., causal relations). To our knowledge, research on scenario and script recognition does not exist apart from the use of symbolic knowledge inspired by the theories of Schank described in Chap. 2.

The police and intelligence domain also shows that we may not be tempted to reduce the content of a text to certain extracted information stored in a template database representation. Sometimes very small details (e.g., a Maria statute on the dashboard of a car) become the key to certain links between the information. In addition, in this domain text is not the only source of information that naturally occurs in unstructured format. Captures of surveillance cameras, images and video cannot be forgotten. Any search and analysis tool will have to deal with these multi-media formats. There is, however, the restriction that many of the information sources are closed sources and are not freely available for training and testing pattern recognizers.

9.6 Information Extraction from Business Texts

The business domain is a domain where structured data go hand in hand with unstructured data, the latter being mostly in the form of text. The text corpora consist of technical documentation, product descriptions, contracts, patents, Web pages, financial and economical news, and increasingly also of informal texts such as blogs and consumer discussions. Data mining has been well established in business communities and explains why mining of texts also becomes highly valued. For these tasks, data and text mining software that is commercially available is often used, offering, however, a very rudimentary solution to the extraction problems. Classical commercial software offers functionality for the clustering of texts, the clustering of terms, the categorization of texts and named entity recognition.

The oldest applications of information extraction technologies are found as part of the processing of *technical documentation* (e.g., for space craft maintenance). In these documents natural language text is interweaved with structured information. Because the documents often have a strict formal organization and follow a number of stylish conventions, their formal characteristics can be fixed and enforced by a drafting tool. Nevertheless, not all content can be structured at drafting time, which leaves room for the extraction of specific information, especially for answering unforeseen questions that users pose.

Businesses are concerned about their *competitive intelligence*. They want to actively monitor available information that is relevant for the decision making process of the company. They can use publicly available sources (e.g., the World Wide Web) in order to detect information on competitors or competitive products by which they might assess the competitive threat. The extracted information can be used in order to react more quickly (e.g., when one of their products has received negative reviews). Information extraction can also be used to find new business opportunities.

Up until now the extraction technologies usually concern *named entity recognition*. Apart from the common named entities such as product brands, organizations and persons, typical business entities can be defined among which are prices and properties of products (e.g., dimensions, availability dates), which can often be seen as attributes of the entities. In this domain one of the earliest relation recognition tasks were developed based on hand crafted symbolic knowledge (e.g., Young and Hayes, 1985). Supervised learning techniques were used in the recognition of company mergers and succession relations in management functions (e.g., Soderland

1999). Unsupervised learning techniques could extract patterns in the financial domain (e.g., Xu et al., 2002).

Problems that can be encountered are that information is often presented in *semi-structured format* (e.g., Web texts where layout characteristics coded in HyperText Markup Language (HTML) play an important role) or business forms with structured information coded in the Extensible Markup Language (XML). The structured characteristics of these documents are often very helpful in order to extract the right information. The problem is that the structured characteristics are *usually not standardized* (e.g., layout or document architectures differ) requiring many annotated texts in order to train suitable systems.

Besides extracting named entities from Web pages, the latest trend is to monitor and analyze the most up to date online news and blog posts, providing immediate insight into consumer discussions, perceptions and issues that could impact brand reputation. Information extraction technology here delivers true market intelligence and provides brand managers, product and marketing professionals with the critical analysis necessary to clearly understand consumer discussions relating to companies, products and competitors. Studies on *sentiment or attitude tracking* are still limited. We refer the interested reader to Hearst (1992), Finn et al. (2002), Dave et al. (2003), Pang and Lee (2004), Riloff et al. (2005), and Shanahan et al. (2006). Sentiment tracking offers a whole new area of research into information extraction where the technologies discussed in this book can be applied.

The business domain will certainly be a large client of extraction technology and offers many opportunities for research. We foresee a growing demand for automated syntheses of information and its presentation (e.g., comparison of prices) that are generated on the basis of flexible information queries. Wrappers that extract information from highly structured sources as the Web have been developed (Kushmerick 2000). The business domain offers a new ground for research in information extraction. When using the World Wide Web as an information source, scalability problems have to be taken care of.

9.7 Information Extraction from Legal Texts

The legal field is perhaps the field where information is almost exclusively found in texts and where huge text collections are available. The repositories of legislation, court decisions and doctrinal texts are increasingly accessible via Web portals. These texts often combine structured with unstructured

data, the former mostly referring to the typical architecture of the documents (e.g., legislation is divided in Books, Chapters, Sections and Articles) or metadata that are typically manually assigned (e.g., date of enactment of an article).

Notwithstanding the large need for information extraction technologies in the legal domain, the use of this technology is very limited. The literature cites the *recognition and tracing of named entities* (mainly persons) (Branting, 2003). The tracing of persons regards the mapping of alias names and the disambiguation of names that have equal writings. Another extraction task regards the *classification of sentences of court decisions according to their rhetorical role* (e.g., offence, opinion, argument, procedure) (Moens et al., 1997; Grover et al., 2003; Hachey and Crover, 2005; Aouladomar, 2005). For the retrieval of court decisions and their use in case based reasoning systems, it is important that the factors, i.e., the fact patterns that play a role in the decision, are assigned to the decision texts and to the arguments of the decisions in particular. The most extensive studies in assigning factors to court decisions were realized by Brüninghaus and Ashley (2001a).

There are different reasons for the low interest of using information extraction techniques in the law domain. A first problem deals with the language of the texts. Legal texts combine ordinary language with a typical legal vocabulary, syntax and semantics, making problems such as disambiguation, part-of-speech tagging and parsing more difficult than would be the case in other texts. Perhaps the most important causes of the slow integration of extraction technologies in the legal domain are a certain resistance to use automated means, the monopoly of a few commercial players that dominate the legal information market, and the past lack of international competitions and golden standards in the form of annotated corpora. In 2006 a legal track of the Text REtrieval Conference is planned.

Nevertheless, comparable to the police domain there is a high demand for extracting named entities such as persons, organizations and locations, for linking them to certain acts or events, and for classifying these factual data into concepts, scripts or issues. The extracted data would be very useful in order to enhance the performance of information retrieval, to perform data mining computations (Stranieri and Zelznikow, 2005) and to facilitate automated reasoning with cases (Brüninghaus and Ashey, 2001b). In addition, information extracted from legislative documents could be integrated in knowledge based systems that automatically answer legal questions.

9.8 Information Extraction from Informal Texts

In all the above cases of information extraction more or less well-formed texts are processed. Often, we are confronted with informal texts from which we want to extract information. Examples are transcribed speech, spam texts, and instant messages that were generated through mobile services. If we can afford to annotate sufficient training examples, simple pattern matching approaches can already be of help. However, in many cases of informal texts the patterns change continuously (e.g., different informal styles of different authors) or deliberately (e.g., in spam mail). Also, the natural language processing techniques on which we rely will not perform as adequate as they should. In the following section we will elaborate on the example of transcribed speech and refer to some other examples of informal texts.

Speech is transcribed to written text by means of automatic speech recognition (ASR) techniques. However, the speech differs from written texts because of the use of different discourse structures and stylistic conventions, and transcribed speech has to cope with the errors of the transcriptions.

Existing information extraction technologies do not perform well on transcribed speech. There are several reasons for this. Orthographic features such as punctuation, capitalization and the presence of non-alphabetic characters are usually absent in transcribed speech. Sentence boundary detection is difficult. Numbers are spelled out in places where digit sequences would be expected. When the vocabulary used by ASR does not contain all entities, detecting unknown names in texts is difficult because orthographic features cannot be used. In most current speech recognition systems, the size and content of the ASR vocabulary is predefined, and the recognizer will output the word in its lexicon that most closely matches the output audio stream. While the overall out-of-vocabulary rate is typically very low (< 1%) for most large-vocabulary systems, the out-of-vocabulary rate is significantly higher for words in proper name phrases, frequently ranging from 5% to more than 20% (Palmer and Ostendorf 2000), and this rate usually differs depending on the type of noun phrase that is considered. The incompleteness of the ASR vocabulary is a common situation in domains where new names constantly surface (e.g., in news and in the business domain).

Most of the research in information extraction from speech regards *named entity recognition*. While recognition of named entities in news stories has attained F-measure values that are comparable with human performance (see *supra*), named entity recognition of speech data – both conversational

speech and broadcast news speech – does not yet attain such a high performance.

The most interesting aspect in the development of information extraction from transcribed speech is the integration of explicit error handling in the extraction system, an idea originally postulated by Grishman (1998). In transcribed speech, errors corrupt the spoken text as words that are wrongly recognized. Consider a simple model in which the errors are created by sending the original text through a noisy channel, which both deletes some incoming tokens and inserts random tokens. Using a pattern recognizer that is trained on a noiseless text will severely reduce the reliability of the information extractor.

Grishman (1998) proposed a symbolic approach characterized by a noisy channel that may insert or delete a noun group or an individual token outside of a noun group. An experiment on MUC-6 texts showed that the model could attain precision values of 85%. But, recall is very low: With a 30% word error, the extraction lost 85% in recall compared to the perfect transcript, meaning that with the model we miss many valid extraction patterns. Papageorgiou et al. (2004) recognized person, location and organization entity names based on a vocabulary name list and a finite-state based named entity recognition module. Although precision values are in the 90%, recall values range between 31% and 53%. The authors blame the lack of proper names in the vocabulary of the ASR engine and the lack of grammar patterns used by the finite state automaton.

Models can be designed that propagate a limited set of alternative analyses from one stage to the next, i.e., from the speech recognition to the extraction. Palmer and Ostendorf (2000) demonstrate the usefulness of this approach. These authors use a hidden Markov model while incorporating part-of-speech features. For the recognition of persons, locations, organizations, timexes and numeric expressions, their model could still attain F-measure rates above 71% for named entity recognition with a word error rate higher than 28% (Evaluation of the DARPA-sponsored Hub-4 Broadcast News Transcription and Understanding).

The use of features different from the typical text based ones is also investigated in information extraction from speech. For instance, features can be considered that mark prosody such as durational, intonational and energy characteristics (e.g., duration of pauses, final vowel, pitch relative to the speaker's baseline). In experiments, prosodic features did not have a notable effect on the F-measure of named entity recognition (Hakkani-Tür et al.,1999).

With the ACE (Automatic Content Recognition) competition we foresee a growing interest in information extraction from transcribed speech as some ACE corpora contain this medium.

In general, informal texts are often ungrammatical. They are characterized by spelling errors, inconsistent use of capitalization patterns, ungrammatical constructions making that simple information extraction tasks such as named entity recognition more difficult while attaining lower accuracy rates than one would expect. Studies on information extraction from informal texts are very limited. Huang et al. (2001) and Jansche and Abney (2002) studied extraction of caller names and phone numbers from voice mail transcripts. Rennie and Jaakkola (2005) extracted named entities from e-mails and bulletin board texts. E-mails often demand inferences of humans for their correct interpretation as content might be left out. The sender and receiver typically share a context, which must be inferred (Minkov et al., 2004). The best that can be done here is taking into account contextual documents. Spam mail is often ungrammatical and the vocabulary is malformed in order to mislead spam filters. Such a situation restricts the application of standard natural language processing tools.

9.9 Conclusions

The case studies demonstrate that information extraction is currently heavily researched. The technologies and algorithms are generically used across domains. While early research primarily relied on symbolic patterns that were manually acquired, current technology is mostly focused towards machine learning of the recognition patterns while handcrafted resources serve as complementary knowledge sources. Current extraction tasks regard named entity recognition, noun phrase coreference resolution (recognition of alias names and disambiguation of polysemous names) and recognition of relations between entity names. In the future we foresee that extraction technologies will be used to build complex scenario's, profiles and scripts and will be integrated in advanced coreferent resolution across documents.

The problems encountered in information extraction are pretty much similar across the different domains. The lack of annotated examples (and corresponding lack of symbolic knowledge when rules are handcrafted) that cover the variety of linguistic expressions is omnipresent. Secondly, the need to find more advanced features to combat certain ambiguities in the patterns learned is also apparent. Increasingly we are confronted with informal texts (e.g., speech, blogs, mails) posing additional challenges on their processing. When dealing with these "noisy" texts, the problems are only cumulated, demanding research in the years to come.

Last but not least, the need for information synthesis is very well present in all applications that attempt to answer complex information ques-

tions. For instance, in news we want to detect and link information about events. In the biomedical domain, we want to automatically discover from texts complex biological scenarios. Police and intelligence services demand to link persons and events in texts allowing them to mine complex criminal patterns. In the business domain, we want to link entities to attribute values such as detailed product information and consumer attitudes. In the legal domain researchers are interested in building complex case representations that will be used in case based reasoning, or in automatically translating legislation into the rules of knowledge based systems that some day might substitute human judges.

In Chap. 7 we have seen that information queries of users are very flexible and that we may not be tempted to represent a document that is used in an information system as a template containing only certain extracted information, but that the extracted information acts as additional descriptors besides the words of the text. These findings form the basis of the last chapter in this book where special attention will go to the role of information extraction in retrieving and synthesizing information.

9.10 Bibliography

Aouladomar, Farida (2005). Some foundational linguistic elements for QA systems: An application to e-government services. In *Proceedings of the Eighteenth JURIX Conference on Legal Knowledge and Information Systems* (pp. 81-90). Amsterdam: IOS Press.

Ashburner, Michael et al. (2000). Gene ontology: Tool for the unification of biology: The Gene ontology consortium. *Nature Genetics*, 25, 25-29.

Bikel, Daniel M., Richard Schwartz and Ralph M. Weischedel (1999). An algorithm that learns what's in a name. *Machine Learning,* 34, 211-231.

Branting, L. Karl (2003). A comparative evaluation of name-matching algorithms. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law* (pp. 224-232). New York: ACM.

Brüninghaus, Stefanie and Kevin, D. Ashley (2001a). Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th International Conference on Artificial Intelligence and Law* (pp. 42-51). New York: ACM.

Brüninghaus, Stefanie and Kevin D. Ashley (2001b). The role of information extraction for textual CBR. In *Proceedings of the 4th International Conference on Case-Based Reasoning – Lecture Notes in Computer Science* (pp. 74-89). Berlin: Springer.

Chau, Michael and Jennifer J. Xu (2005). CrimeNet explorer: A framework for criminal network knowledge discovery. *ACM Transactions on Information Systems,* 23 (2), 201-226.

Crystal, Michael R. and Carrie Pine (2005). Automated org-chart generation from intelligence message traffic. In *Proceedings of the 2005 International Conference on Intelligence Analysis.*

Cullota, Aron and Jeffrey Sorensen (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 424-430). East Stroudsburg, PA: ACL.

Dave, Kushal, Steve Lawrence and David M. Pennock (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the Twelfth International World Wide Web Conference.* New York: ACM.

Dowman, Mike, Valentin Tablan, Cristian Ursu, Hamish Cunningham and Borislav Popov (2005). Semantically enhanced television news through Web and video integration. In *Proceedings of the World Wide Web Conference.* New York: ACM.

Finkel, Jenny, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex and Claire Grover (2005). Reporting the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics 2005,* 6 (Suppl I): S5.

Finn, Aidin, Nicholas Kushmerick and Barry Smyth (2002). Genre classification and domain transfer information filtering. In Fabio Crestani, Mark Girolami and Cornelis J. van Rijsbergen (Eds.), *Proceedings of ECIR-2, 24th European Colloquium on Information Retrieval Research.* Heidelberg: Springer.

Friedman, Carol, Pauline Kra, Hong Yu, Michael Krauthammer and Andrey Rzetsky (2001). GENIES: A natural language processing system for the extraction of molecular pathways from journal articles. *ISMB* (*Supplement of Bioinformatics*), 74-82.

Gaizauskas, Robert J., George Demetriou and Kevin Humphreys (2000). Term recognition and classification in biological science journal articles. In *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP* (pp. 37-44).

Glenisson, Patrick, Janick Mathijs, Yves Moreau and Bart De Moor (2003). Meta-clustering of gene expression data and literature-extracted information. *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 5 (2), 101-112.

Grefenstette, Gregory, Yan Qu, James G. Shanahan and David A. Evans (2004). Coupling niche browsers and affect analysis for an opinion mining. In *Proceedings RIAO 2004.* Paris: Centre des Hautes Études.

Gooi, Chung Heong and James Allan (2004). Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 9-16). East Stroudsburgh, PA: Association for Computational Linguistics.

Grishman, Ralph (1998). Information extraction and speech recognition. In *Proceedings of the Broadcast News Transcription and Understanding Workshop* (pp. 159-165).

Grishman, Ralph, Silja Huttunen and Roman Yangarber (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics,* 35, 236-246.

Grover, Claire, Ben Hachey, Ian Hughson and Chris Korycinski (2003). Automatic summarization of legal documents. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law* (pp. 243-251). ACM: New York.

Hakkani-Tür, Dilek, Gökhan Tür, Andreas Stolcke and Elizabeth Shriberg (1999). Combining words and prosody for information extraction from speech. In *Proceedings EUROSPEECH '99, 6th European Conference on Speech Communication and Technology*.

Hearst, Marti (1992). Direction-based text interpretation as an information access refinement. In Paul Jacobs (Ed.), *Text-Based Intelligent Systems*. Hillsdale, NJ: Lawrence Erlbaum.

Huang, Jing, Geoffry Zweig and Mukund Padmanabhan (2001). Information extraction from voicemail. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 290-297). San Mateo, CA: Morgan Kaufmann.

Huang, Minlie et al. (2004). Discovering patterns to extract protein-protein interactions from full text. *Bioinformatics,* 20 (18), 3604-3612.

Jansche, Martin and Steven P. Abney (2002). Information extraction from voicemail transcripts. In *Proceedings of Empirical Methods in Natural Language Processing.* East Stroudsburg, PA: ACL.

Kanehisa Minoru, Susumu Goto, Shuichi Kawashima, Yasushi Okuno and Masahiro Hattori (2004). The KEGG resource for deciphering the genome. *Nucleic Acid Res,* 32, D277-280.

Kushmerick, Nicholas (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118, 15-68.

Kou, Zhenzhen, William W. Cohen and Robert F. Murphy (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics*, Suppl1, i266-io273.

Lee, Ki-Joong, Young-Sook Hwang, Seonho Kim and Hae-Chang Rim (2004). Biomedical named entity recognition using a two-phase model based on SVMs. *Journal of Biomedical Informatics*, 37, 436-447.

Leroy, Gondy, Hinchun Chen and Jesse D. Martinez (2003). A shallow parser based on closed-class words to capsule relations in biomedical text. *Journal of Biomedical Informatics*, 36, 145-158.

Li, Xin, Paul Morie and Dan Roth (2004). Robust reading: Identification and tracing of ambiguous names. In *Proceedings of the Human Language Technology Conference* (pp. 17-14). East Stroudsburgh, PA: ACL.

Marcotte, Edward M., Ioannis Xenarios and David Eisenberg (2001). Mining literature for protein-protein interactions. *Bioinformatics*, 17 (4), 359-363.

Minkov, Einat, Richard C. Wang and William W. Cohen (2004). Extracting personal names from emails. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)* (pp. 443-450). East Stroudsburg, PA: ACL.

Moens, Marie-Francine, Caroline Uyttendaele and Jos Dumortier (1997). Abstracting of legal cases: The SALOMON experience In *Proceedings of the 6th International Conference on Artificial Intelligence and Law* (pp. 114-122). New York: ACM.

Ng, Vincent and Claire Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 104-111). San Francisco, CA: Morgan Kaufmann.

Ng, Vincent and Claire Cardie (2003). Weakly supervised natural language learning without redundant views. In *Proceedings of the Human Language Technology Conference* (pp. 183-180). East Stroudsburgh, PA: ACL.

Palmer, David D. and Mari Ostendorf (2000). Improving information extraction by modelling errors in speech recognizer output. http://citeseer.ist.psu.edu/646402.html

Pang, Bo and Lilian Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 115-124). East Stroudsburg, PA: ACL.

Papageorgiou, Harris, Prokopis Prokopidis, Iason Demiros, Nikos Hatzigeorgiou and George Carayannis (2004). CIMWOS: A multimedia retrieval system based on combined text, speech and image processing. In *Proceedings of the RIAO 2004 Conference.* Paris: Centre des Hautes Études.

Ramani, Arun K., Razvan C. Bunescu, Raymond J. Mooney and Edward M. Marcotte (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology,* 6: R40.

Rennie, Jason D.M. and Tommie Jaakkola (2005). Using term informativeness for named entity detection. In *Proceedings of the Twenty-Eight Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 353-360). New York: ACM.

Riloff, Ellen, Janyce Wiebe and William Phillips (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence* (*AAAI-05*). Menlo Park, CA: AAAI Press.

Shanahan, James G., Yan Qu and Janyce Wiebe (Eds.) (2006). *Computing Attitude and Affect in Text* (*The Information Retrieval Series* 20). New York: Springer.

Soderland, Stephen (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning,* 1-3, 233-272.

Stapley, BJ, Kelley LA and Sternberg MJ (2002). Predicting the sub-cellular location of proteins from using support vector machines. *Pacific Symposium Biocomputing*, 374-385.

Stranieri, Andrew and John Zeleznikow (2005). *Data Mining in Law.* New York: Springer.

Xu, Feiyu, Daniela Kurz, Jacub Piskorski and Sven Schmeier (2002). Term extraction and mining of term relations from unrestricted texts in the financial domain. In *Proceedings of the 5th International Conference on Business Information Systems BIS-2002* (pp. 304-310). Poznan, Poland.

Young, Sheryl R. and Philip J. Hayes (1985). Automatic classification and summarization of banking telexes. In *The Second Conference on Artificial Intelligence* (pp. 402-408). Los Alamitos, CA: IEEE Press.

Zhang, Jie, Dan Shen, Guodong Zu, Jian Su and Chew-Lim Tan (2004). Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37, 411-422.