Semantic Tools for Forensics: Approaches in Forensic Text Analysis

Michael Spranger and Dirk Labudde
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*name.surname*}@hs-mittweida.de

Abstract—**The analysis of digital media and particularly texts acquired in the context of police securing/seizure is currently a very time-consuming, error-prone and largely manual process. Nevertheless, such analysis are often crucial for finding evidential information in criminal proceedings in general as well as fulfilling any judicial investigation mandate. Therefore, an integrated computational solution for supporting the analysis and subsequent evaluation process is currently developed by the authors. In this work, we present an approach for categorizing texts with adjustable precision combining rule-based decision formula and machine learning techniques. Furthermore, we introduce a text processing pipeline for deep analysis of forensic texts as well as an approach for the identification of criminological roles.**

Keywords*—*forensic; ontology; German; text processing

I. Introduction

The analysis of texts that are subject of legal considerations with the goal of obtaining criminalistic evidence is a branch of general linguistics [1]. Such texts are retrieved by persons involved in the criminal proceedings from a variety of sources, e.g., secured or confiscated storage devices, computers and social networks. Forensic texts, as considered in this work, relate to textual data that may contain evidential information. In contrast to the texts usually considered in scientific work focussing text processing tasks this kind of texts are neither clearly defined nor thematically unified. Additionally, such texts may vary in quality with respect to their grammar, wording and spelling which strongly depends on the author's language skills and the target audience. Rather, textual data of different type and origin need to be meaningfully linked to answer a specific criminological question reasonably and above all accurately. Furthermore, forensic linguistics cover beside other research topics, utterance and word meaning or authorship analysis and proof [2].

The results of these analyses are used to solve other more complex problems in the criminal investigations, like

- recognition and separation of texts with a case-related criminalistic relevance
- recognition of relations in these texts in order to reveal whole relationship networks and planned activities
- identification and/or tracking of fragmented texts
- identification or tracking of hidden semantics

In the considered context, the term *hidden semantics* is synonymous with one kind of linguistic steganography. In this work only the first two points are in the focus. However, this kind of deep analysis takes a long time, especially if the amount and heterogeneity of data, the fast changeover of communication forms and communication technologies is taken into account. In order to solve this problem, computer linguistic methods and technologies can be applied. These are originated in the crossover of linguistics and computer sciences [3]. The complexity of the evaluation makes it difficult to develop one single tool covering all fields of application. In order to address this problem, a domain framework is currently under development (see [4] for further discussions).

As a consequence of the analysis of the secured data from a historical case of business crime and the exploration of the special needs of criminologists discussed in Section II we present in this work a pipeline for categorizing texts with adjustable precision using an approach which is combined of rule-based decision formula and machine learning techniques. Especially that leaves the opportunity to the criminologist to decide whether the specificity (precision) is more important or the sensitivity (recall), although a high sensitivity may be of greater practical importance. Thus, a high sensitivity is principally necessary to find all incriminating or even exculpatory documents but the results need to be filtered manually since they may be interspersed with irrelevant documents, whereas a high specificity is sometimes more appropriate to get a quick overview about the corpus. Furthermore, we outline a text processing pipeline for deep analysis of forensic texts based on these insights and a rule-based approach for identifying special roles of named entities. Currently, the text categorization module is evaluated in practice whereas the deep analysis pipeline including the role identification is under implementation.

In the next Section the peculiarities of the considered kind of texts is shown at a glance. Subsequently, a pipeline for analysing forensic texts deeply as well as a first approach for detecting forensic roles is outlined before a practicable method for categorizing such texts is introduced and discussed.

II. Assessment of Requirements

This work focusses textual data secured by police officers as part of the evidence process. Hence, for the purposes of this work historical data in a case of business crime is provided by the prosecutorial. A first manual assessment of these data enables to determine, whether:

- the data material is of considerable heterogeneity related to its structure and domain

- important information may be situated in non-text based data (e.g photocopies of invoices)
- there are totally irrelevant texts that may hide relevant information through their abundance (e.g forms, templates)
- information may have been deliberately obscured in order to protect them from discovery
- some texts can be characterized by strong syntactic weaknesses
- some texts may be fragmented by erasing/reconstruction

These specific characteristics distinguish the examined corpus from other corpora commonly used and evaluated in research.

Further, a survey made by the authors, which was conducted by affiliated criminologists has revealed that finding and separating relevant documents seized in the database is the most time consuming and difficult part during the evaluation.

III. Approaches in Forensic Text Analysis

In this section, several strategies for handling forensic texts respecting the insights from the needs assessment (section II) are introduced. Since the most aspects of this work are currently under implementation no final results will be presented yet. Thus, these aspects are outlined subsequently.

A. Pipeline for Deep Analysis

The deep analysis of forensic texts has to respect their characteristics described in the previous section. It includes particularly tasks in Information/Event Extraction to instantiate a criminological ontology as the central element in the solution developed under this work. In particular, the work of Wimalasuriya and Dou [5], Embley [6] and Maedche [7], shows that the use of ontologies is suitable for assisting the extraction of semantic units as well as their visualization and structures such processes very well. We have divided the whole process in three sub-processes:

1) creation of both the criminological ontology and the analysis corpus
2) basic textual processing and detection of secondary contexts
3) instantiation of the ontology and iteratively refinement

In order to define the extraction tasks as well as to introduce case-based knowledge the first of all is the creation of the criminological ontology in its specialized form as Topic Map we have developed in an earlier work [4]. This step may be supported by using existing ontologies created in similar previous cases. Subsequently, the analysis corpus needs to be created, especially for separating the textual data from other files and extracting the raw texts from the documents including optical character recognition in cases of digital images like photocopies. This data is stored in a database together with extracted meta-data and added to an index for quick access. In the second step some state-of-the-art textual processing steps like Part-of-Speech-tagging, language recognition and some special operations for structured texts may be performed. Especially, we detect event-narrative documents. This task has been introduced by Huang and Riloff [8] for exploring secondary contexts. They define these as sentences that are not explicitly part of the main event description. Nevertheless, these secondary contexts could yield information related to the event of interest that could provide important evidence or lead to the booty, further victims or accomplices. The final step within the main process is constituted by the actual extraction process. Here, the actual event sentences that are suitable to instantiate at least one part of the ontology are recognized and, if needed, extracted together with the information from secondary contexts. Then, we try to refine the instantiated model iteratively by identifying forensic roles as described in III-B. Figure 1 illustrates the whole process schematically.

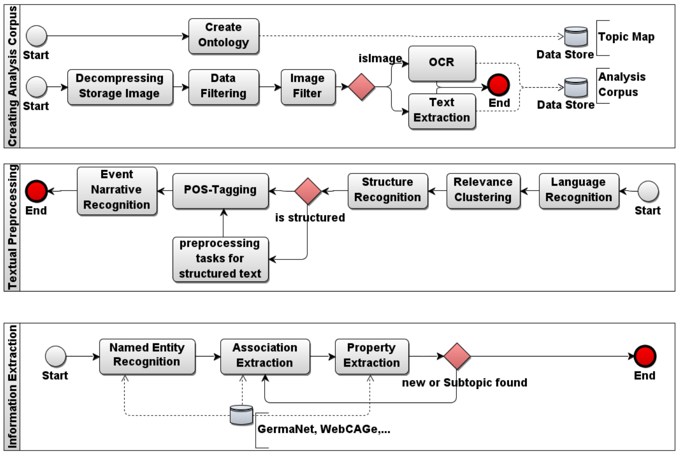

Fig. 1. The tool-pipeline for deep analysis. We have divided the whole process in three sub-processes: 1) creating analysis corpus 2) textual preprocessing 3) information extraction

B. Identification of Forensic Roles

The recognition of named entities is a well-researched part of Text Mining and a regular task in every Information/Event Extraction solution as well as in our pipeline mentioned in III-A. The general task is to identify all instances $i \in I$ of each concept $c \in C$ taking into account their hypernymy and hyponymy relationships. This task can be solved practically by using Gazeteer-based solutions via supervised learning methods [9], [10] up to the usage of semi-/unsupervised learning approaches [11]. However, no existing solution we applied has been proven itself to be able to assign forensic roles. The assignment of such a role is often dependent on more than one document as well as the contribution of case-based knowledge by the criminologist. Therefore, our framework is based on an ontology acting as an extraction and visualization template that is able to provide such knowledge. The ontology model we used is based on the Topic Map standard. In our previous work [4] we stated that each topic can contain a set of facets. These facets are used beside others to model rules that an inference machine can use to reason the appropriate role of an entity within a post-process. In this way the level of detail within the computational recognition of entities is able to be increased. Figure 2 shows a detail of a fictional forensic Topic Map that may have been created by a criminologist. Here, a accomplice is described as a person that satisfies one or two of the following rules:

- the person has common interest in the deed exactly when he has instantiated an association possess with the topic booty
- the person has shared worked exactly when their related instance in the Topic Map has an instantiated association drive to an instance of the topic getaway-car

The number of rules that have to be satisfied depends on rule weights which act as indicators for rule importance. The concrete instance defines the same facets with binary values depending on the matching behaviour of each rule.

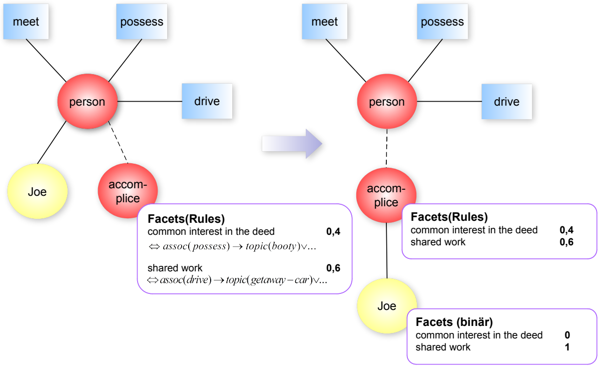
meet
possess
person
drive
Joe
accom-plice
Facets(Rules)
common interest in the deed 0,4
$\Leftrightarrow assoc(possess) \rightarrow topic(booty) \vee ...$
shared work 0,6
$\Leftrightarrow assoc(drive) \rightarrow topic(getaway-car) \vee ...$
meet
possess
person
drive
accom-plice
Facets(Rules)
common interest in the deed 0,4
shared work 0,6
Joe
Facets (binär)
common interest in the deed 0
shared work 1

Fig. 2. Gradually refining of named entities. The entity *Joe* as instance (yellow circle) of the abstract topic (red circle) *person* can gradually assigned to their concrete manifestation *accomplice* which is a subtopic by iterative comparison of its facets lodged as rules.

C. Categorization of Forensic Texts

As discussed in Section II, filtering and categorization is the most important task in evaluation of forensic texts and a regular Information Retrieval task. Categorization as a specialization of classification aims to place a document in one small set of categories using machine learning techniques. More formal, given a set of documents $D = \{d_1, ..., d_m\}$ and further a set of categories $C = \{c_1, ..., c_n\}$ the task can be described as an surjective mapping $f : C \rightarrow D$. Ikonomakis et al. [12] have given an overview about supervised machine learning methods for solving this problem. However, they observed that the performance is significantly depending on a corpus of high quality and sufficient size. Riloff and Lehnert [13] introduced an approach for high-precision text classification. The augmented relevancy signature algorithm they introduced reached up to 100% precision with over 60% recall on the MUC-4 corpus. Nevertheless, in the focussed domain these results are not always sufficient especially since they do not relate to the properties of forensic texts. It has to be emphasized, that each false-negative (a not identified, case-relevant document) could provide crucial evidences. This highlights the necessity for a method which yields 100% recall with justifiable precision. Beebe and Clark [14] has introduced an approach to handle the information overload resulting from the recall-precision trade-off problem. They considered a similar problem and suggest to cluster the results thematically. However, designing and training a suitable classifier is a challenging problem. Since the knowledge of the criminologist (general and case-based) is available related to a concrete judicial investigation order, rules can improve the performance in some cases. This leads to a combined approach. Since the categories has modelled as a taxonomy tree we can extend this model so that we are able to assign a set of rules (e.g., regular expressions applied on the documents body) to each category. These rules are combined by disjunction within the categories itself and by conjunction between different categories in cases of one continuous chain of parent-child relationships (figure 3). Each of these rules has to define the target that it should applied on (e.g., file name or content), a rule type that helps to select the corresponding rule solver and the rule itself. In this way, we are able to select a certain number of seeds that ensure high precision which is required to start an appropriate bootstrapping machine learning algorithm to classify the remaining documents (figure 4). Notice, the performance can be influenced by rephrasing the corresponding rules, since the performance of a bootstrapping algorithm significantly depends on the seed elements chosen, more precise their representativeness. Thus, strictly formulated rules may result in high precision but low recall, whereas applying more weak rules will increase the recall. First measures of performance using probability-based classifiers, like Naive Bayes, as well as similarity-based classifiers, like k-NN or TF-IDF shows that the performance reaches up to

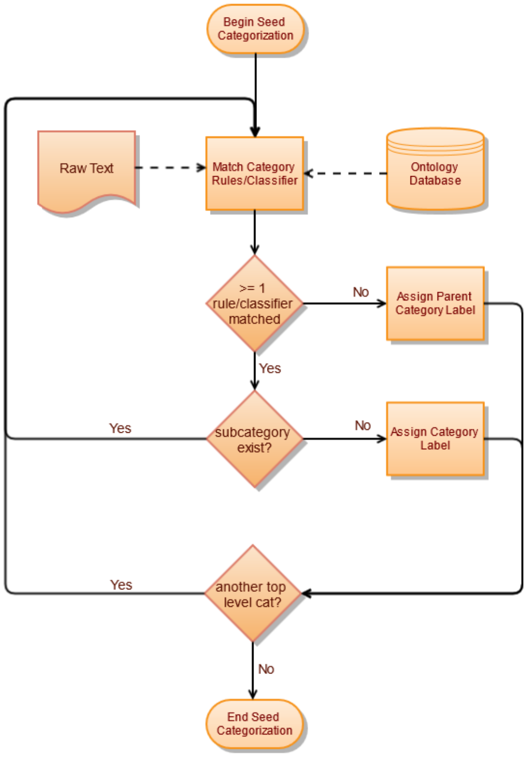

Fig. 3. *Acquisition of seed documents:* The raw text under consideration is checked against a set of category rules recursively. Starting at a top-level category, at least one category rule/classifier has to match until the match of each subcategory, drawn from recursion, has failed. In this way only the label of the most specific category starting at each existing top-level category is assigned.

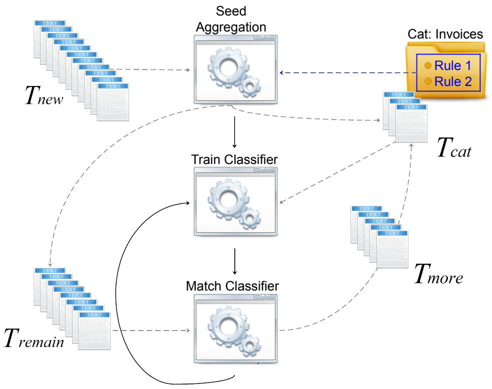

Fig. 4. Bootstrapping Algorithm for classifying forensic texts. From the texts T_{new} a set of seed documents for each category is acquired using the rules annotated in the taxonomy. This set T_{cat} is used to train one initial weak binary classifier per category. Subsequently, this classifier is used to classify the remaining texts T_{remain} and store the new labelled documents T_{more} to T_{cat}. Finally, the classifier is going to be improved iteratively using T_{cat} until no document is left or no further improvement is possible.

100% precision and recall applied on the corpus provided by the prosecutorial as mentioned in Section II depending on the employed algorithm and the concrete category. This result could be a consequence of classifier over-fitting caused by the underlying homogeneous corpus. We have observed that in the in the corpus we used the documents are characterized by great similarity. Therefore, a more appropriate corpus is created currently.

One of the biggest advantages of this combined approach lays in the adjustable precision depending on an intelligent combination of rules and machine learning algorithms.

IV. Conclusion

In this work, we have outlined some kernel processes for information extraction in the environment of the criminal proceedings. These processes are suitable to deal with very heterogeneous data concerning their domain as well as their quality. In the task of deep exploration of the raw data there was great emphasis on the discovery of all relevant information using secondary contexts to avoid misunderstandings and lacks in the evidence. In the identification of forensic roles we have described a new approach in refining ontology instances by deriving and applying semantic roles logic-based. A corresponding module using Prolog is currently under development. In the task of classification of forensic texts we have to respect that each misclassified file could lead to a lack of evidence. Therefore, it must be ensured that at best no type II errors occur during the categorization. At the same time the taxonomy definition has to remain flexible. Because of a lack of training data supervised learning is not applicable. Therefore, a bootstrapping approach is chosen, combined with a rule-based search for seed files we have earned very good preliminary results at 100% accuracy in selected domains. However, this unexpected result could be due to an over-fitting to the used corpus. For this reason we currently create a new extended Corpus with the support of the prosecutorial.

Acknowledgment

The authors would like to thank the members of the criminal investigation department and the prosecutorial Chemnitz (Germany). We acknowledge funding by ”Europäischer Sozialfonds” (ESF), the Free State of Saxony and the University of Applied Sciences Mittweida.

References

[1] H. Kniffka, Working in Language and Law. A German perspective. Palgrave, 2007.

[2] E. Fobbe, Forensische Linguistik - Eine Einführung. (Forensic Linguistics - An Introduction) Narr Frankcke Attempto Verlag, 2011.

[3] K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer, Computerlinguistik und Sprachtechnologie - Eine Einführung (Computational Linguistics and Language Technology - An Introduction), 3rd ed. Spektrum Akademischer Verlag, 2010.

[4] M. Spranger, S. Schildbach, F. Heinke, S. Grunert, and D. Labudde, “Semantic tools for forensics: A highly adaptable framework,” in Proc. 2nd. International Conference on Advances in Information Management and Mining, IARIA. ThinkMind Library, 2012, pp. 27 – 31.

[5] D. C. Wimalasuriya and D. Dou, “Ontology-based information extraction: An introduction and a survey of current approaches,” Journal of Information Science, vol. 36, no. 3, 2010, pp. 306–323.

[6] D. W. Embley, “Toward semantic understanding: an approach based on information extraction ontologies,” in Proceedings of the 15th Australasian database conference - Volume 27, ser. ADC ’04. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2004, pp. 3–12.

[7] A. Maedche, G. Neumann, and S. Staab, “Bootstrapping an ontology-based information extraction system,” Studies In Fuzziness And Soft Computing, vol. 111, 2003, pp. 345–362.

[8] R. Huang and E. Riloff, “Peeling back the layers: detecting event role fillers in secondary contexts,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, ser. HLT ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1137–1147.

[9] A. Mikheev, M. Moens, and C. Grover, “Named entity recognition without gazetteers,” in Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, ser. EACL ’99. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp. 1–8.

[10] G. Zhou and J. Su, “Named entity recognition using an hmm-based chunk tagger,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ser. ACL ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 473–480.

[11] Z. Kozareva, “Bootstrapping named entity recognition with automatically generated gazetteer lists,” in Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, ser. EACL ’06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 15–21.

[12] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, “Text classification using machine learning techniques,” WSEAS Transaction on Computers, vol. 4, no. 8, 2005, pp. 966–974.

[13] E. Riloff and W. Lehnert, “Information extraction as a basis for high-precision text classification,” Transactions on Information Systems, vol. 12, no. 3, 1994, pp. 296–333.

[14] N. L. Beebe and J. G. Clark, “Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results,” Digital Investigation, vol. 4, 2007, pp. 49–54.