Towards Establishing an Expert System for Forensic Text Analysis

Michael Spranger and Dirk Labudde
Department of Mathematics, Physics, Informatics
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*michael.spranger, dirk.labudde*}@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 and knowledge-based computational solution for supporting the analysis and subsequent evaluation process is currently developed by the authors. In this work, we outline the main ideas of this framework and 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 approaches towards solving domain specific problems like detection and understanding of hidden semantics as well as the automatic assignment of forensic roles.**

***Keywords*–forensic, ontology, German, text processing, expert system, text analysis, Topic Map, categorization, classification**

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], [2]. 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 criminalistic question reasonably and above all accurately. Furthermore, forensic linguistics cover beside other research topics, utterance and word meaning or authorship analysis and proof [3].

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 but not restricted to this. Rather, even the use of slang afflicted language let known text mining algorithms fail. Understanding hidden semantics is one of the hardest tasks during the analysis of forensic texts not only for machines but also for humans. Among other things, for this reason, 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 [4]. 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 [5] 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 that is a combination of rule-based decision formula and machine learning techniques. Especially, that leaves the opportunity to the criminologist to decide whether the specificity (precision) or the sensitivity (recall) is more important. 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. Subsequently, we introduce two approaches towards solving the hidden semantics problem. Currently, the text categorization module is evaluated in practice whereas the deep analysis pipeline including the role identification as well as the hidden semantics detector is under implementation.

In the next Section, the peculiarities of the considered kind of texts is shown at a glance. In Section III, a special crime ontology acting as foundation model for an expert system in the field of forensic text analysis is presented. Subsequently, in Section IV a pipeline for analysing forensic texts deeply as well as a first approaches for detecting forensic roles and hidden semantics 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 persons involved in the criminal investigations 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 local 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 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 criminalists, has revealed that finding and separating relevant documents seized in the database is the most time consuming and difficult part during the evaluation.

III. Development of a Crime Ontology

A. Ontology-based Information Extraction

The term *ontology* is commonly understood as a formal and explicit specification of a common conceptualization. In particular, it defines common classified terms and symbols referred to a syntax and a network of associate relations [6], [7]. Developing ontologies for criminalistic purposes is a prior condition for annotating texts and raise questions in this particular domain. The term *taxonomy* as a subset of ontology is used for the classification of terms (concepts) in ontologies and documents. On the one hand, a criminalistic ontology is characterised by its case-based polymorphic structure and on the other hand by special terms used in criminal proceedings. This aspect has to be taken into account by the definition of any ontology representation model, as we see in Section III-C. Ontologies can be divided into two levels of generality. A *domain ontology* models the knowledge of an almost highly-specialised domain as a part of the real world in an extensive and profound manner. An *upper ontology* describes the common objects applicable to a wide range of domain ontologies. Furthermore, it creates a glossary of basic terms and object descriptions used in various relevant domains [7].

Cowie and Wilks [8] constitute Information Extraction (IE) as a process for selectively structuring and combining data, located, explicitly stated or implied in various texts. A slightly more formal view is given by Russell and Norvig. They understand IE as the acquisition of knowledge by searching occurrences of objects of specific classes and relations between them within natural language text [9].

The process of IE can be supported by ontologies in several ways. The usage as *extraction ontology* is one way to participate in the benefits of ontologies. In this case the IE process itself is guided by using templates generally used by sophisticated techniques of knowledge representation [7], [10]. Presenting the output of the IE process using ontologies is another way supporting this process.

Combining both approaches we obtain an IE system that is supported at most by ontologies. Such systems are called Ontology-based Information Extraction (OBIE)-systems [10].

B. Representation of Knowledge Models

The representation of ontologies can be realized through different models with different levels of expressiveness. Taxonomies and thesauri, which are not mentioned here, can be considered as simple ontologies under the adherence to certain conventions. Instead of this, some more expressive models will be introduced in this section.

The intention of concept maps, as developed by Josef Novak at the Cornell University [11], is to represent relationships between concepts. According to this, a concept map is an abstract description of certain ideas or of a specific knowledge domain. They visualize semantic units (prepositions) for a certain domain, while semantic units consist of two terms (concepts) connected through a named relation. Labelling a relation provides a higher degree of understanding through additional semantic information. It is explicitly not forbidden to create cross relations between multiple concepts [7], [11].

Topic map is the most expressive model and well defined because of its ISO standardisation. There is a wide variety of implementations, e.g., XML Topic Maps (XTM), transposing the basic concepts of this standard although they ignore or modify single aspects defined by the ISO standard.

The standard *ISO/IEC 13250* describes the usage of topic maps in the areas of information exchange, organization and representation with the aid of topics. Basically, structural information provided by topic maps allow to describe relations between topics, related to abstract things, and to attach addressable information objects to a single topic (occurrences). The nature of all constituent parts can be described more in detail by using properties (facets). Another significant point is that the information objects used in a topic map can be assigned to a scope as described in more detail in Section III-C. It is important to know that several topic maps can provide structural information referring to the same resource. In this way, the architecture enables the combination of topic maps and the coupling of information from different areas. Because of their extrinsic character topic maps can be seen as an extension or overlay of information objects. In summary it can be stated that topic maps enable versatile and simultaneous views at information objects, whose structural nature is principally unrestricted. Hence, it is possible to use an object-oriented, hierarchical, sorted or unsorted approach or each combination

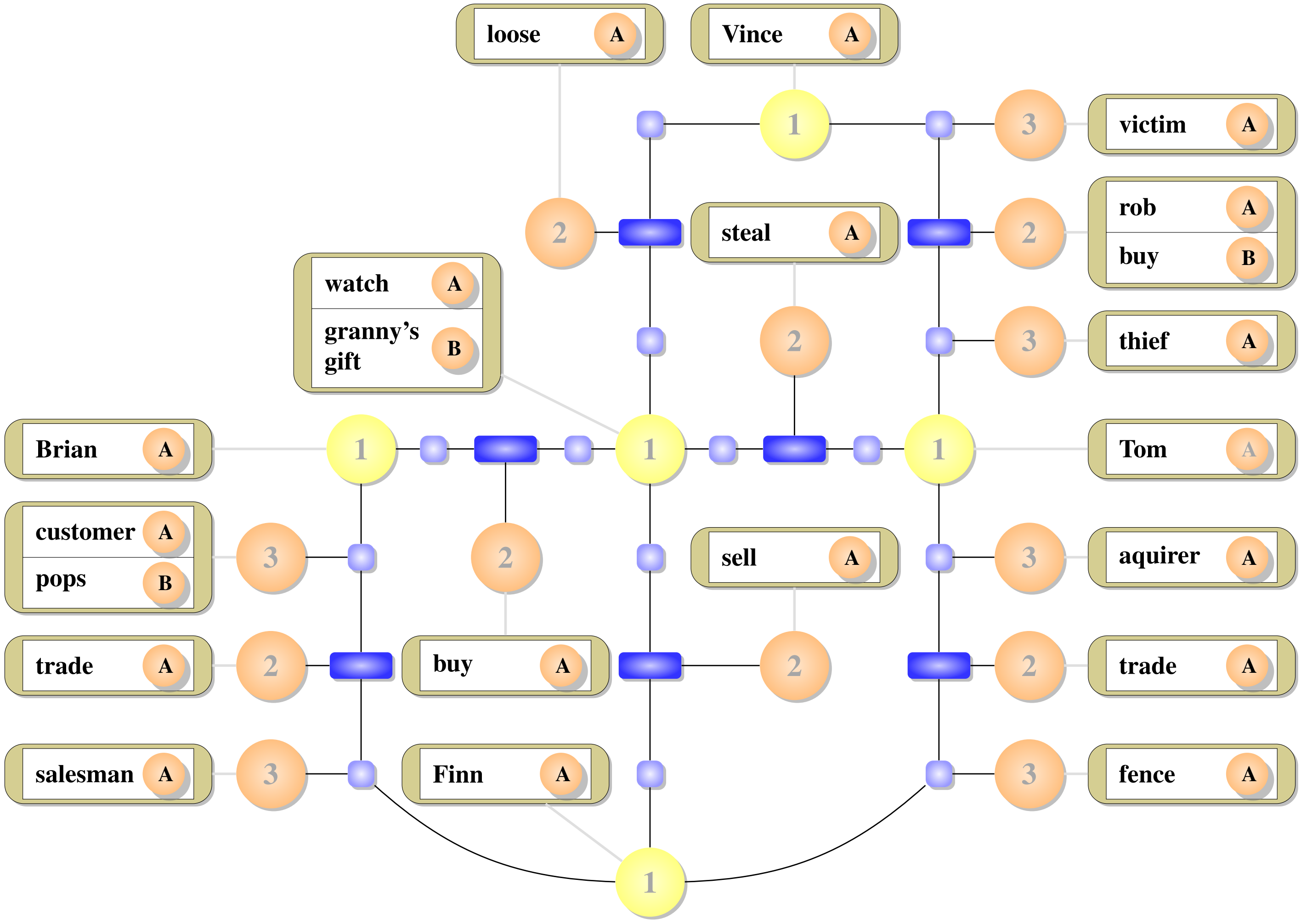


Figure 1: Extract of an ontology used for the description of property crimes. It demonstrates a typical interaction of the different topic map elements, whereas familiar relations are not included here.

of these. Additionally, it is possible to overlay an unrestricted count of topic maps on a given set of information resources [12].

C. Crime Ontology Model

In this project we use a modified variant of the topic map standard to model an ontology, where the created model is based on the contents and thoughts of the ISO standard without claiming a full implementation of all parts. In general, major semantic elements can be considered to be present in the model while most syntactic elements have been replaced with elements as required by a model driven software development. Especially, the use of *scopes* within topic maps is a significant advantage for modelling multilingualism and improves the determination of meanings. In the field of crime sciences and forensic linguistics multilingualism is not only restricted to native and foreign languages, moreover it is possible to integrate slang afflicted language groups, dialects and different verbal skills. Furthermore, *scopes* offer one possibility to solve the hidden semantics problem, we considered in Section I, by annotating one or more different meanings directly to the particular topic. The topic map elements used in the model considered here are described in Table I.

Figure 1 demonstrates an application of the topic map derivative as developed under this work for modelling a criminalistic ontology – a simple case of uncovering a ring dealing with stolen goods. The shown extract does not cover all elements of the topic map model implemented.

The core objects in the example network are highlighted by the number *1* – the persons Vince, Tom, Finn and Brian, as well as the item watch. Associations specified through descriptive topics between these objects are highlighted with the number *2*. A specified role, taken by an object within an association, is highlighted with the number *3*.

Taking a closer look at the example shown in Figure 1 leads to the suggestion that the course of creating this network could have been happened the following way: Brian is searching for a watch because his old one is broken. He asks in different stores for a model fitting his needs till he finds a salesman (Finn) who offers him that he might get one in his next delivery. A few days later Finn calls Brian that he got a watch for him, Brian does not hesitate and buys it. After a closer examination at home he comes over a nearly faded

TABLE I: Elements of the forensic Topic Map model

Element	Description
Subject (Topic)	an abstract or concrete entity in the domain to be analysed
Instance (Topic)	the concrete manifestation of a subject (*red circle*)
Descriptor (Topic)	typifies any other syntactical elements (*orange circle*); i.e., adds further details related
Association	a relation between two topics, usually subject and instance (*light blue rhomboid*)
Association Role	specifies the roles of the topics in an association (*blue square*)
Occurrence	corresponds to the crete manifestation of a topic in a resource, usually related to an Instance.
Topic Name	is the name representation of topics (*green rounded rectangle*)
Name Item	denotes the name of a specific topic, associated to a Scope (*white rectangle within the topic name*)
Facet	names a class of attributes of a topic and can include several Facet Values
Facet Value	a particular attribute as distinct value; can be a topic or another Facet
Scope	defines semantic layers; e.g., causing system to focus by filtering particular syntactical elements

inscription on the back of the watch and shows it his friend, a policeman. Some days earlier the policeman was called by Vince, a person who lost his heirloom at the beach, which has an inscription just like this watch. They went back to the store together where Finn was spotted by the policeman, known to him from smaller complaints by different customers. After some consideration time the police confiscated Finns computer. Within the analysis of the confiscated material an instant messaging protocol reveals the following snippet:
Tom: "I bought granny's gift, which pops demanded."
Finn: "Alright, bring it over."
Where Tom is also known to the police with no familiar relations to Finn. Some further background work reveals the full potential of their relation and completes the network. Reconsidering all the facts Finn can be marked as a fence who sells stolen goods acquired by Tom. He kept looking for a watch described by Brian and finally found a model easy to steal, Vince's watch. Lucky coincidence in this posed example for demonstrating the cooperation of the different elements of the ontology model to uncover a fence network.

IV. 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 only 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 [10], Embley [13] and Maedche [14], 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, which we have developed in an earlier work [5]. 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 also 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 [15] 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 IV-B. Figure 2 illustrates the whole process schematically.

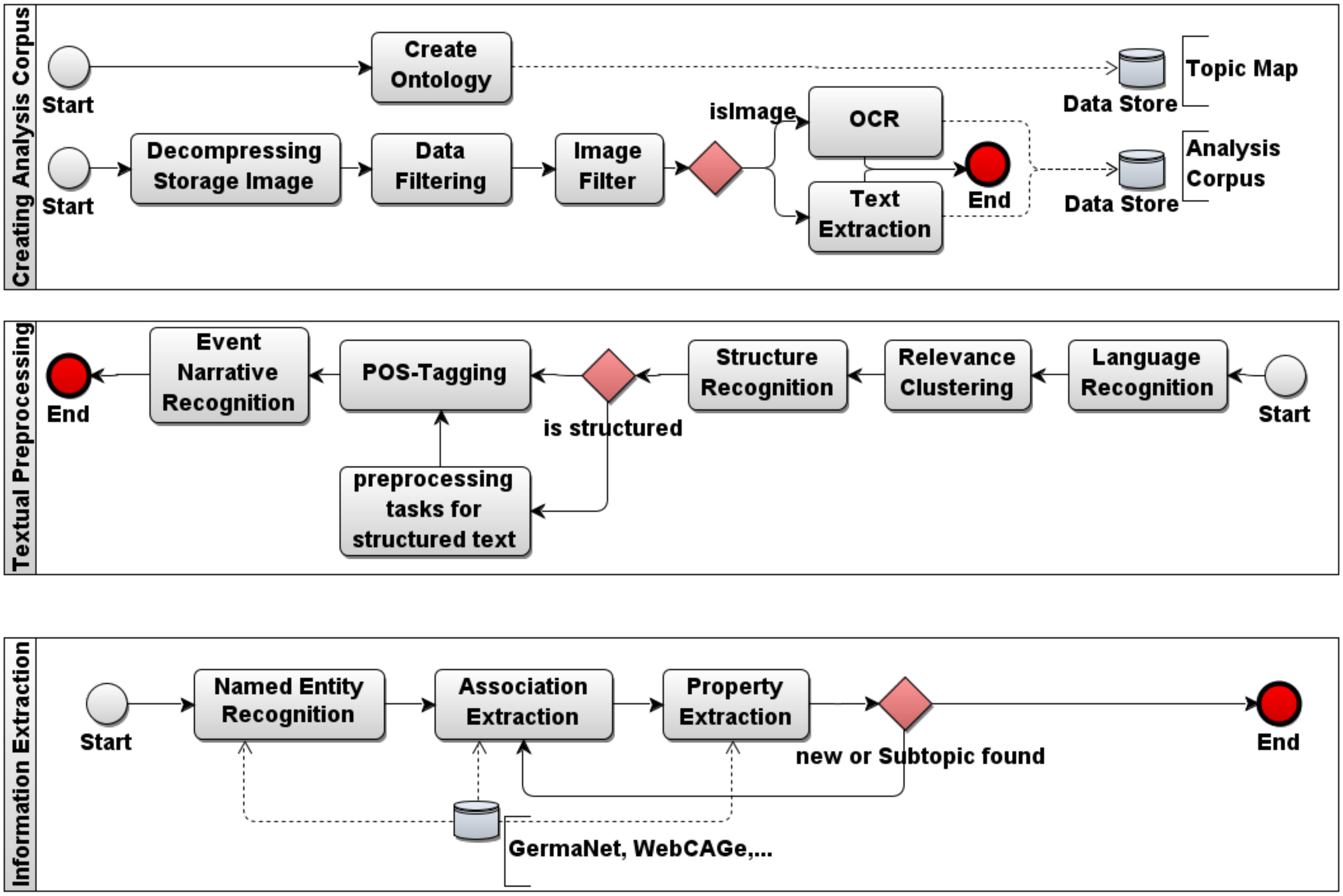


Figure 2: The tool-pipeline for deep analysis. We have divided the whole process in three sub-processes: 1) creating analysis corpus 2) textual preprocessing 3) information extraction

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 Section IV-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 [16], [17] up to the usage of semi-/unsupervised learning approaches [18]. However, no existing solution we applied has been proven itself to be able to assign forensic roles. The assignment of such a role often depends on more than one document as well as on the contribution of case-based knowledge by the criminalist. 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 [5], 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 3 shows a detail of a fictional forensic Topic Map that could have been created by a criminalist. 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 instance of a topic acting in the role of 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 car acting in the role of a means of escape

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.

C. *Towards Solving the Hidden Semantics Problem*

As mentioned in Section I the hidden semantics problem is one of the hardest tasks during the analysis of forensic texts even for criminalists or linguistic experts with years of experience. Thus, this problem can only be solved by consideration of the whole context and the knowledge of experts. A system that should be able to detect or even solve this problem automatically needs to process the overall IE-tasks before. Since knowledge extracted automatically as well

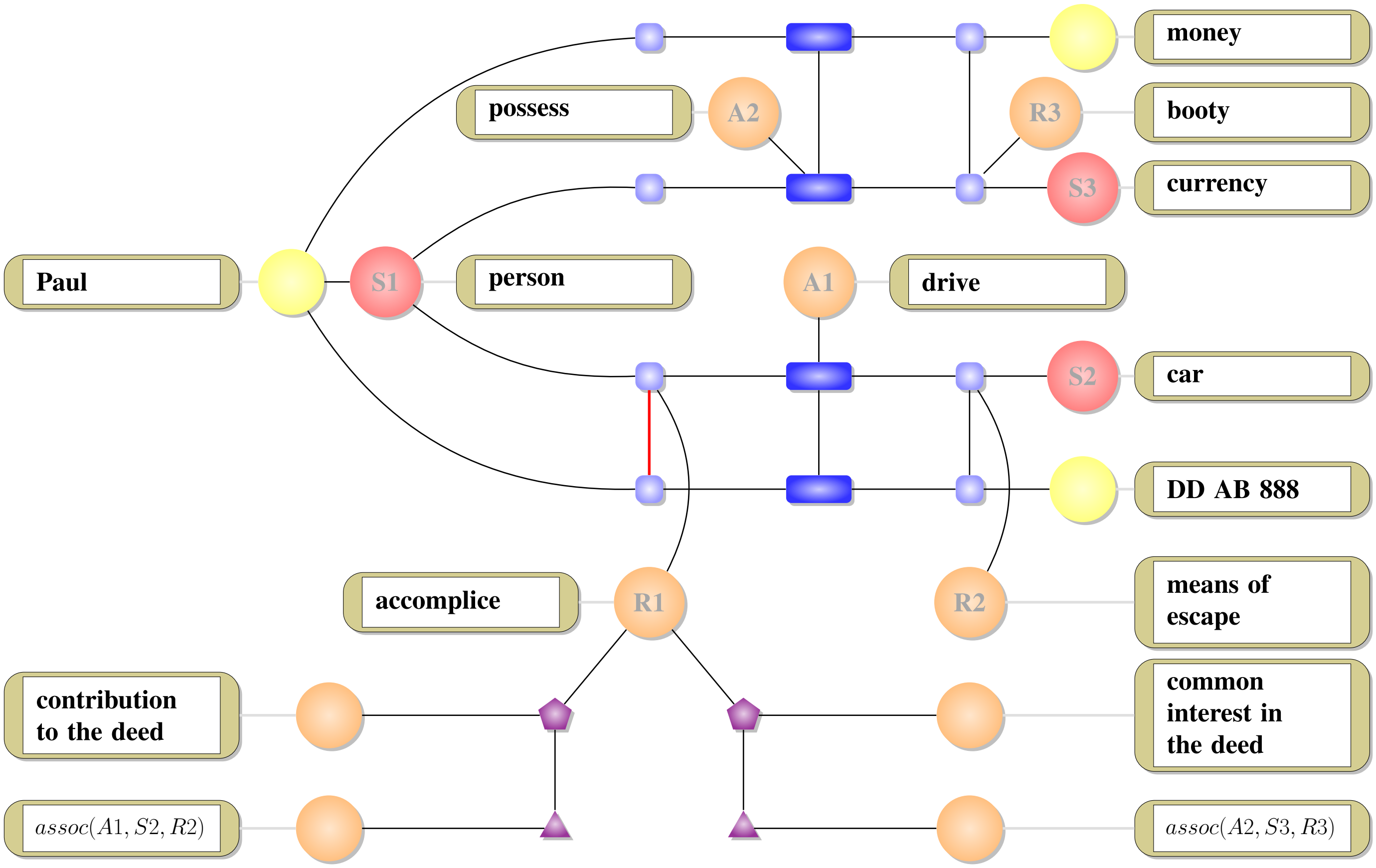


Figure 3: Gradually refining of named entities. The entity *Paul* as instance (yellow circle) of the abstract topic (red circle) *person* can be gradually assigned to their concrete manifestation *accomplice*, which is a subtopic by iterative comparison of its facets lodged as rules.

as introduced by experts is represented by a criminalistic Topic Map (see Section III-C), hidden semantics might be detected by considering its special features. Maicher has introduced an approach for merging Topics with the same meaning modelled by different authors in an distributed world [19]. This leads to a similar approach for the problem discussed here. Thus, each instance a system may find is clearly defined by the position of the related topic within the taxonomy, its facets and the set of instantiated associations where it plays a highly specific role. We assume this semantic context will remain approximately constant if the text is transposed towards a steganographic code, because only the wording changes (see Figure 4a).

More formal, let each Instance $i \in I$ be well defined by a tupel $\{T, F_T, R_A, A_T\}$, where T is the related Topic-hierarchy, F_T is a set of Facets of each of this Topics that discriminates the instance from other similar ones, R_A is a set of Roles that it plays relating to a set of Associations and finally A_T is a set of Associations of each Topic. This tupel constitutes the context $C(i)$ of a specific Instance. Subsequently, each context has to be compared with the context of other Topics using a distance function $dist$ to find out the degree of similarity. The definition of a threshold ϵ supports the decision, whether two topics are possibly the same or not [see equations (1) and (2)].

$$\Delta_{min}(C(i)) = min_{j \in I \setminus i}\{dist(C(i), C(j))\} \qquad (1)$$

$$SYN(C(i), C(j) = \begin{cases} 1, & C(j)\ has\ \Delta_{min}(C(i)) < \epsilon \\ 0, & else \end{cases} \qquad (2)$$

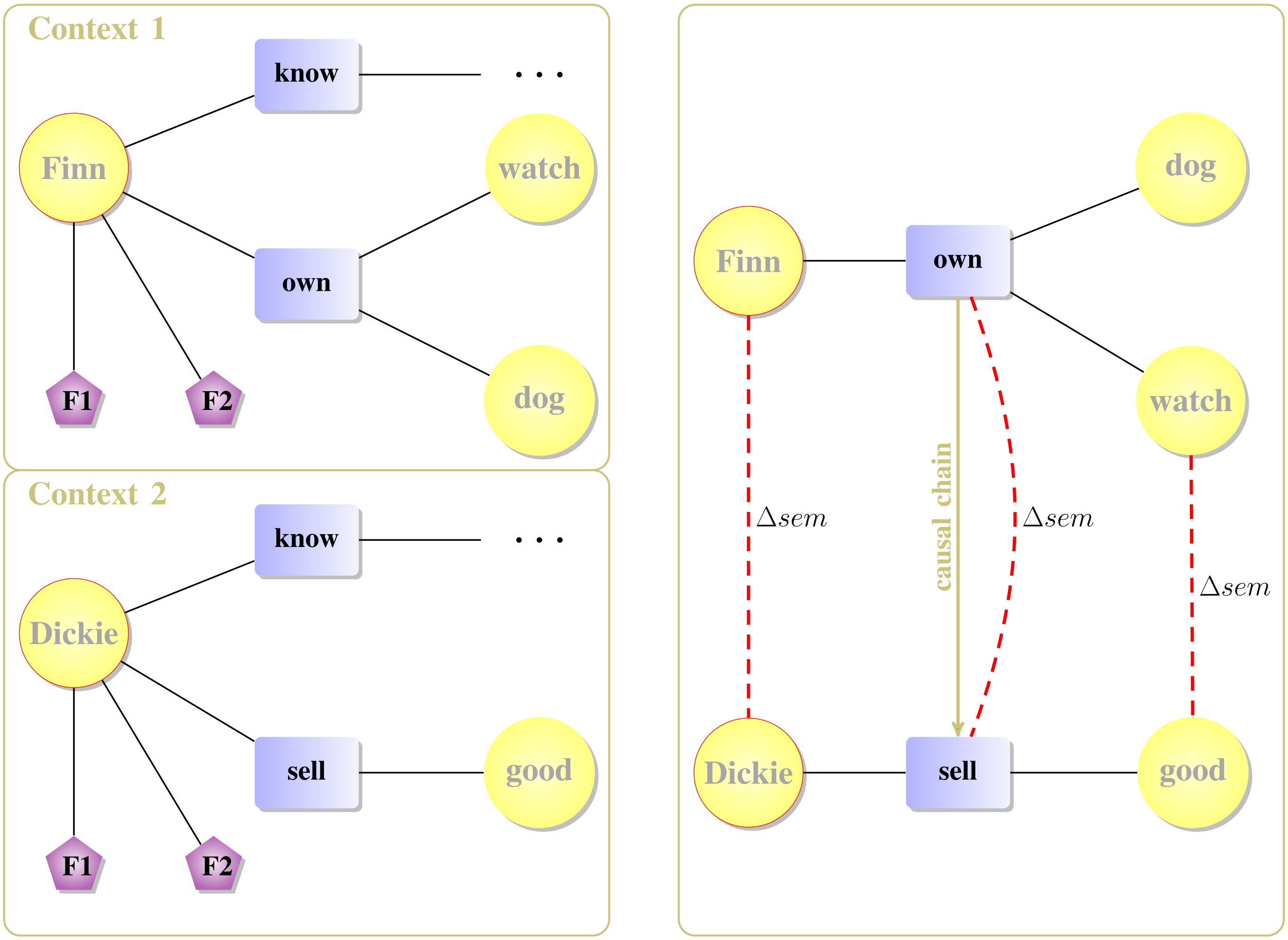
In order to determine the distance between contexts the semantics in the ontology need to be encoded in a numeric format. For Topics the method of Wang et al. [20] can be adapted, whereby the farther away from one Topic to another, the less similarity is determined by the constant k [see equation (3)]. This constant needs to be determined empirically.

$$S_T(t) = \begin{cases} 1, & t = T \\ max\{k * S_T(t') \mid t' \in children(t)\}, & else \end{cases} \qquad (3)$$

Another approach is more Association-centred. We consider alignments of all Associations within the same causal chain and calculate an edit distance. This distance measure is related to distances in the ontology-graph (see Figure 4b). Formally, let A be the set of associations and K the set of causal chains that may be derived from A. A causal chain is constituted by all associations $\{a_1 ... a_n\}$, whereby $a_1 \rightarrow a_2 \rightarrow ... \rightarrow a_n$. Further, let A_T be the set of Associations related to an specific Topic. The causal chains in that we are interested in can be described as

$$K_{relevant} = \{k \in K \mid \exists a, b \in k \land a \in A_{T1} \land b \in A_{T2}\} \qquad (4)$$

Let S be the set of sentences that can be built using any association in one k. Thus, we can calculate a score for each



(a) Topic-centred approach - Assuming the two instances *Finn* and *Dickie* referring to the same entity. Each of these defines a context that contains the facets, associations and roles derived from the topic they are associated with. The level of similarity between the two contexts also indicates the similarity of the instances.

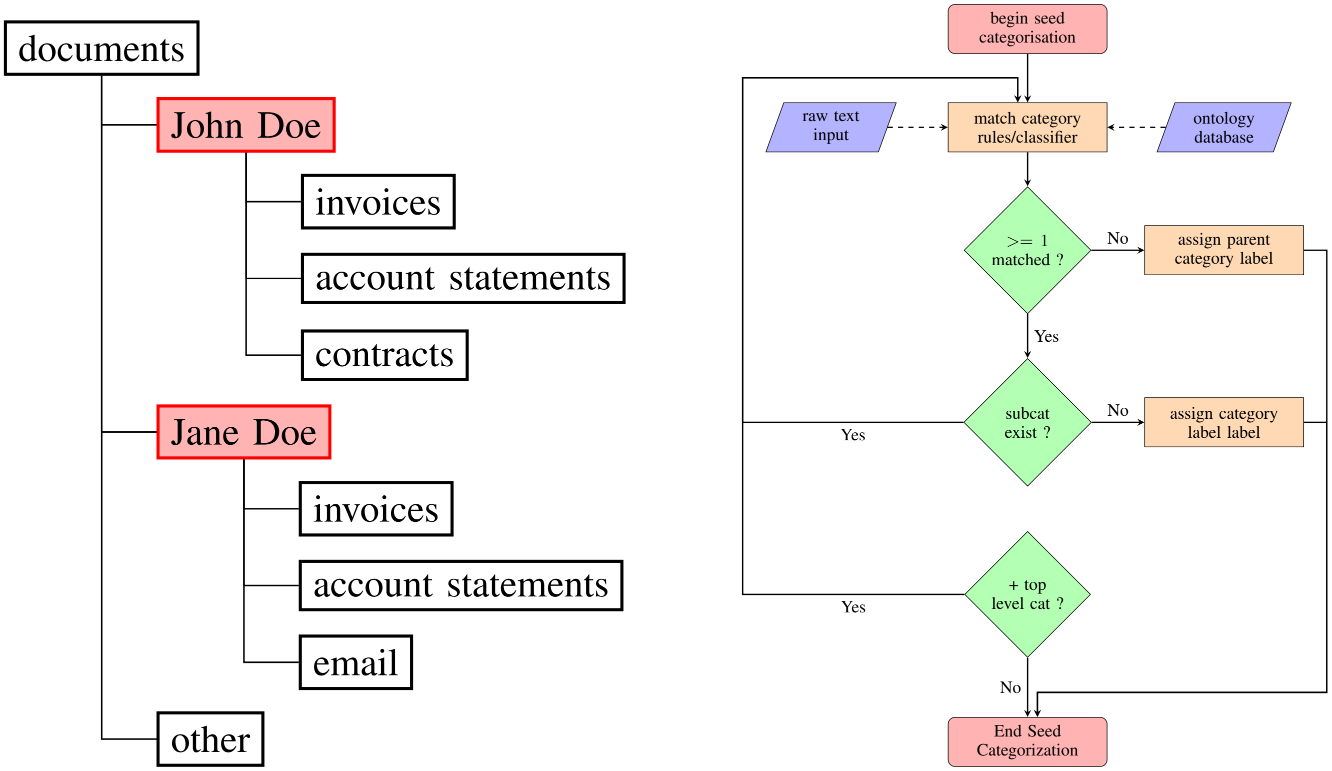
(b) Association-centred approach - Assuming the two instances *Finn* and *Dickie* are the same. The similarity of the two instances can be determined by pairwise alignment of all possible sentences and calculating a semantic distance. Since the associations *own* and *sell* are situated in one causal chain the probability for synonymity of the instances will increase.

Figure 4: Detection of Hidden Semantics

alignment $\{(a, b) \mid a, b \in S\}$. The higher this score the higher the probability that the Topics involved have the same meaning.

D. 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. [21] 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 [22] 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 [23]. 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 that yields at best 100% in sensitivity with justifiable precision. Beebe and Clark [24] have introduced an approach to handle the information overload resulting from the sensitivity-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. Due to the fact that the knowledge of the criminalist (general and case-based) is available related to a concrete judicial investigation order, rules can improve the performance in some cases. Since the categories are modelled as a taxonomy tree we can extend this model so that we are able to assign a set of rules (e.g., regular



(a)

(b)

Figure 5: *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.

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 5a). 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 6). The whole selection process of seed documents is shown in Figure 5b. Notice, the performance of the machine learning algorithm used 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 sensitivity, whereas applying more weak rules will increase the sensitivity.

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 100% precision with 93.58% sensitivity applied on the corpus provided by the prosecutorial as mentioned in Section II. The results are depending on the employed algorithm and the concrete category and 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 remarkable similarity. Therefore, a more appropriate corpus is created currently. For lack of an additional real-life corpus we cross-checked our results using a subset of the 20-Newsgroups-Corpus [25] consisting of the categories *med* and *space*. Depending on the chosen start-rules we achieved sensitivity between 87.6% and 92.4% with precision between 52.3% and 100% (F1 66.79% - 93.39%). This result confirms the strong dependence of the rules used. 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.

V. 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 we put 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 the logic programming

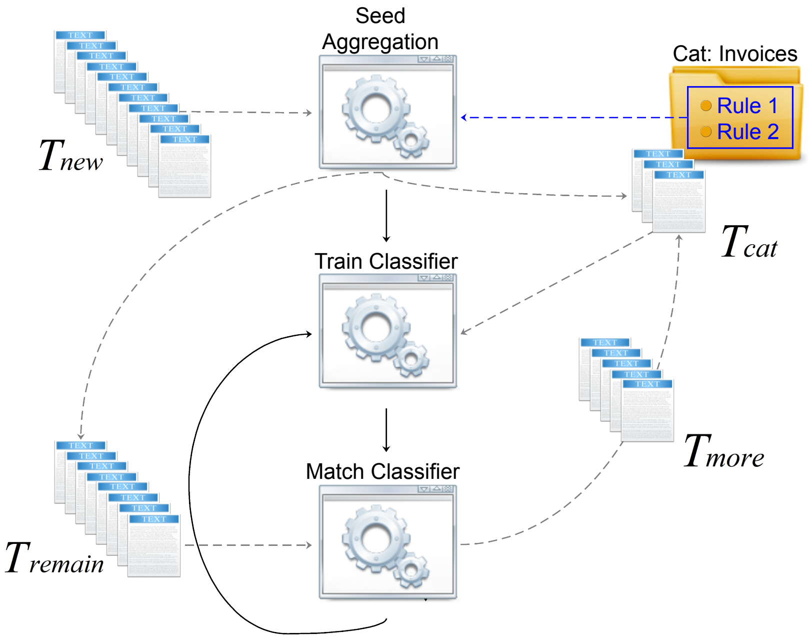


Figure 6: Bootstrapping algorithm for classifying forensic texts. From the texts T_{new} a set of seed documents for each category is acquired using the rules annotated in the taxonomy. This set T_{cat} is used to train one initial weak binary classifier per category. Subsequently, this classifier is 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.

language *Prolog* is currently under development. Furthermore, we introduced two approaches towards solving the hidden semantic problem. Both are based on the calculation of a semantic distance measure using the forensic topic map model we presented at the very beginning. 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 up to 100% precision with 93.58% (F1 = 96.68%) sensitivity in selected domains. However, this unexpected result could be due to an over-fitting to the used corpus. For this reason we currently creating a new extended corpus with the support of the local 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] M. Spranger and D. Labudde, “Semantic tools for forensics: Approaches in forensic text analysis,” in Proc. 3rd. International Conference on Advances in Information Management and Mining (IMMM), IARIA. ThinkMind Library, 2013, pp. 97–100.

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

[3] E. Fobbe, Forensische Linguistik - Eine Einführung. Narr Frankcke Attempto Verlag, 2011.

[4] K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer, Computerlinguistik und Sprachtechnologie - Eine Einführung, 3rd ed. Spektrum Akademischer Verlag, 2010.

[5] 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 (IMMM), IARIA. ThinkMind Library, 2012, pp. 27–31.

[6] T. R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing,” in Formal Ontology in Conceptual Analysis and Knowledge Representation, N. Guarino and R. Poli, Eds. Kluwer Academic Publishers, 1993.

[7] A. Dengel, Ed., Semantische Technologien, 1st ed. Spektrum Akademischer Verlag, 2012.

[8] J. Cowie and Y. Wilks, “Information extraction,” in Handbook of Natural Language Processing., H. M. R. Dale and H. Somers, Eds. New York: Marcel Dekker, 2000.

[9] S. Russell and P. Norvig, Künstliche Intelligenz: Ein moderner Ansatz, 3rd ed. Paearson Deutschland, 2012.

[10] 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.

[11] J. D. Novak and D. B. Gowin, Learning how to learn. Cambridge University Press, 1984.

[12] ISO/IEC, Topic Maps, Information Technology, Document Description and Processing Languages, ISO/IEC Std. ISO/IEC 13 250, Rev. Second Edition, 2002.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] L. Maicher, “The impact of semantic handshakes,” in Proceedings of the 2nd international conference on Topic maps research and applications, ser. TMRA’06. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 140–151.

[20] J. Z. Wang, Z. Du, R. Payattakool, S. Y. Philip, and C.-F. Chen, “A new method to measure the semantic similarity of go terms,” Bioinformatics, vol. 23, no. 10, 2007, pp. 1274–1281.

[21] 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.

[22] 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.

[23] “MUC Data Sets,” 2014, URL: http://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html [accessed: 2014-01-06].

[24] 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.

[25] “20 Newsgroups,” 2014, URL: http://qwone.com/ jason/20Newsgroups/ [accessed: 2014-01-06].