Meta-Level Information Extraction

Peter Kluegl, Martin Atzmueller, and Frank Puppe

University of Würzburg,
Department of Computer Science VI
Am Hubland, 97074 Würzburg, Germany
{pkluegl, atzmueller, puppe}@informatik.uni-wuerzburg.de

Abstract. This paper presents a novel approach for meta-level information extraction (IE). The common IE process model is extended by utilizing transfer knowledge and meta-features that are created according to already extracted information. We present two real-world case studies demonstrating the applicability and benefit of the approach and directly show how the proposed method improves the accuracy of the applied information extraction technique.

1 Introduction

While structured data is readily available for information and knowledge extraction, unstructured information, for example, obtained from a collection of text documents cannot be directly utilized for such purposes. Since there is significantly more unstructured (textual) information than structured information e.g., obtained by structured data acquisition information extraction (IE) methods are rather important. This can also be observed by monitoring the latest developments concerning IE architectures, for example *UIMA* [1] and the respective methods, e.g., conditional random fields (CRF), support vector machines (SVM), and other (adaptive) IE methods, cf., [2–4].

Before IE is applied, first an IE model is learned and generated in the learning phase. Then, the IE process model follows a standard approach depicted in Figure 1: As the first step of the process itself, the applicable features are extracted from the document. In some cases, a (limited form of) knowledge engineering is used for tuning the relevant features of the domain. Finally, the generated model is applied on the data such that the respective IE method selects or classifies the relevant text fragments and extracts the necessary information.

With respect to the learning phase, usually machine-learning related approaches like candidate classification, windowing, and markov models are used. Often SVMs or CRFs are applied for obtaining the models. The advantages of CRFs are their relation to the sequence labeling problem, while they do not suffer from the dependencies between the features. However, a good feature selection step is still rather important. The advantages of SVMs are given by their automatic ranking of the input features and their ability to handle a large number of features. Therefore, less knowledge engineering is necessary. However, the IE task can also be implemented by knowledge engineering approaches applying rules or lambda expressions. In the case studies we will present a rule-based approach that is quite effective compared to the standard approaches.

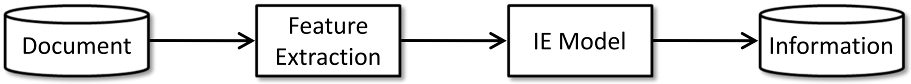


Fig. 1. Common Process Model for Information Extraction.

Specifically, this paper proposes extensions to the common IE approach, such that meta-level features that are generated during the process can be utilized: The creation of the meta-level features is based on the availability of already extracted information that is applied in a feedback loop. Then repetitive information like structural repetitions can be processed further by utilizing transfer knowledge. Assuming that a document, for example, is written by a single author, then it is probably the case that the same writing and layout style is used for all equivalent structures. We present two case studies demonstrating the applicability and benefit of the approach and show how the proposed method improves the accuracy of the applied information extraction technique.

The rest of the paper is structured as follows: Section 2 presents the proposed novel process model for information extraction extending the standard process. We first motivate the concrete problem setting before we discuss the extensions in detail and specifically the techniques for meta-level information extraction. After that, Section 3 presents two real-world case studies: We demonstrate the applicability of the presented approach, for which the results directly indicate its benefit. Next, we discuss related work. Finally, Section 4 concludes with a summary of the presented approach and points at interesting directions for future research.

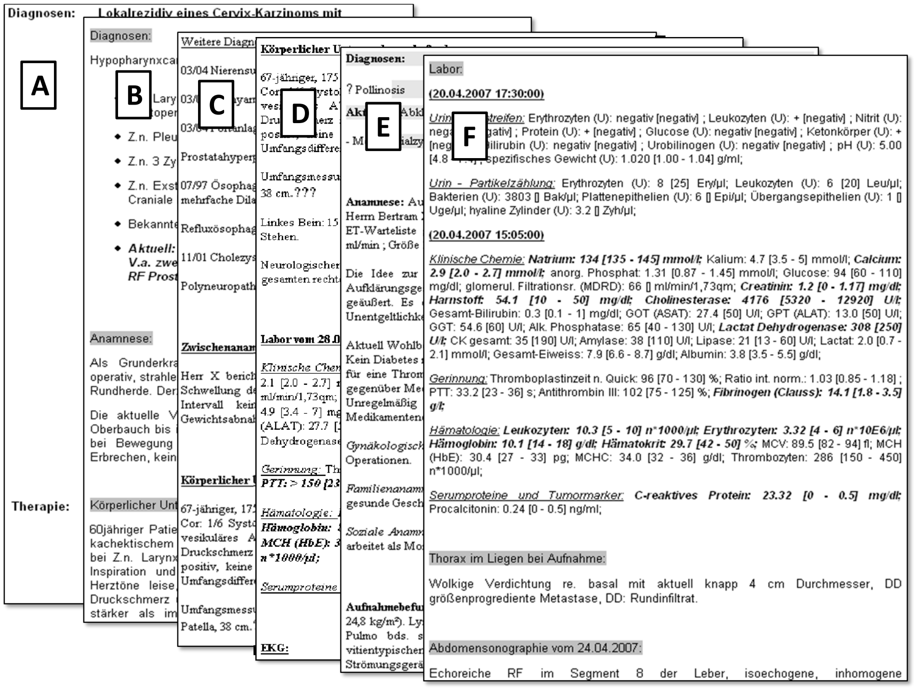
2 Meta-Level Information Extraction

In the following, we first motivate the proposed approach by presenting two examples for which the commonly applied standard process model is rather problematic. Next, we present the process model for meta-level information extraction and discuss its elements and extensions in detail.

2.1 Problem Statement

To point out certain flaws of the standard process, we discuss examples concerning information extraction from curricula vitae (CV) and from medical discharge letters. Both examples indicate certain problems of the common process model for information extraction and lead to following claim:

> Using already extracted information for further information extraction can often account for missing or ambiguous features and increase the accuracy in domains with repetitive structure(s).



Labor:

(20.04.2007 17:30:00)

Urin streifen: Erythrozyten (U): negativ [negativ] ; Leukozyten (U): + [negativ] ; Nitrit (U): nega gativ] ; Protein (U): + [negativ] ; Glucose (U): negativ [negativ] ; Ketonkörper (U): + [neg Bilirubin (U): negativ [negativ] ; Urobilinogen (U): negativ [negativ] ; pH (U): 5.00 [4.8 ; spezifisches Gewicht (U): 1.020 [1.00 - 1.04] g/ml;

Urin - Partikelzählung: Erythrozyten (U): 8 [25] Ery/µl; Leukozyten (U): 6 [20] Leu/µl; Bakterien (U): 3803 [] Bak/µl; Plattenepithelien (U): 6 [] Epi/µl; Übergangsepithelien (U): 1 [] Uge/µl; hyaline Zylinder (U): 3.2 [] Zyh/µl;

(20.04.2007 15:05:00)

Klinische Chemie: ***Natrium: 134 [135 - 145] mmol/l;*** Kalium: 4.7 [3.5 - 5] mmol/l; ***Calcium: 2.9 [2.0 - 2.7] mmol/l;*** anorg. Phosphat: 1.31 [0.87 - 1.45] mmol/l; Glucose: 94 [60 - 110] mg/dl; glomerul. Filtrationsr. (MDRD): 66 [] ml/min/1,73qm; ***Creatinin: 1.2 [0 - 1.17] mg/dl; Harnstoff: 54.1 [10 - 50] mg/dl; Cholinesterase: 4176 [5320 - 12920] U/l;*** Gesamt-Bilirubin: 0.3 [0.1 - 1] mg/dl; GOT (ASAT): 27.4 [50] U/l; GPT (ALAT): 13.0 [50] U/l; GGT: 54.6 [60] U/l; Alk. Phosphatase: 65 [40 - 130] U/l; ***Lactat Dehydrogenase: 308 [250] U/l;*** CK gesamt: 35 [190] U/l; Amylase: 38 [110] U/l; Lipase: 21 [13 - 60] U/l; Lactat: 2.0 [0.7 - 2.1] mmol/l; Gesamt-Eiweiss: 7.9 [6.6 - 8.7] g/dl; Albumin: 3.8 [3.5 - 5.5] g/dl;

Gerinnung: Thromboplastinzeit n. Quick: 96 [70 - 130] %; Ratio int. norm.: 1.03 [0.85 - 1.18] ; PTT: 33.2 [23 - 36] s; Antithrombin III: 102 [75 - 125] %; ***Fibrinogen (Clauss): 14.1 [1.8 - 3.5] g/l;***

Hämatologie: ***Leukozyten: 10.3 [5 - 10] n*1000/µl; Erythrozyten: 3.32 [4 - 6] n*10E6/µl; Hämoglobin: 10.1 [14 - 18] g/dl; Hämatokrit: 29.7 [42 - 50] %;*** MCV: 89.5 [82 - 94] fl; MCH (HbE): 30.4 [27 - 33] pg; MCHC: 34.0 [32 - 36] g/dl; Thrombozyten: 286 [150 - 450] n*1000/µl;

Serumproteine und Tumormarker: ***C-reaktives Protein: 23.32 [0 - 0.5] mg/dl;*** Procalcitonin: 0.24 [0 - 0.5] ng/ml;

Thorax im Liegen bei Aufnahme:

Wolkige Verdichtung re. basal mit aktuell knapp 4 cm Durchmesser, DD größenprogrediente Metastase, DD: Rundinfiltrat.

Abdomensonographie vom 24.04.2007:

Echoreiche RF im Segment 8 der Leber, isoechogene, inhomogene

Fig. 2. Examples of different headlines in medical discharge letters.

CVs For the extraction from CV documents, a predefined template with slots for start time, end time, title, company and description is filled with the corresponding text fragments. The text segments describing experiences or projects are used to identify a template. Then, the slots of the templates are extracted. Often the company can be identified using simple features, e.g., common suffixes, lists of known organizations or locations. Yet, these word lists cannot be exhaustive, and are often limited for efficency reasons, e.g., for different countries. This can reduce the accuracy of the IE model, e.g., if the employee had been working in another country for some time. Humans solve these problems of missing features, respectively of unknown company names, by transfering already 'extracted' patterns and relations. If the company, for example, was found in the third line of ninety percent of all project sections, then it is highly probable that an 'unclear' section contains a company name in the same position.

Medical discharge letters Medical discharge letters contain, for example, the observations, the history, and the diagnoses of the patient. For IE, different sections of the letter need to be identified: The headlines of a section cannot only help to create a segmentation of the letter, but also provide hints what kind of sections and observations are present. Since there are no restrictions, there is a variety of layout structure; Figure 2 shows some examples: Whereas the headlines in (A) are represented using a table, (C+D) use bold and underlined. However, (B) and (F) color the headlines' background and use bold and underlined for subheaders. Some physicians apply layout features only to emphasize results and not for indicating a headline. It is obvious, that a classification model faces problems, if the relation between features and information differs for each input document. In contrast, humans are able to identify common headlines using the containing words. Then, they transfer these significant features to other text fragments and extract headlines with a similar layout.

2.2 Process Model

The human behaviour solving the flaws of the common IE process model seems straight forward, yet its formalization using rules or statistical models is quite complex. We approach this challenge by proposing an extended process model, shown in figure 3: Similar to the common IE process model, features are extracted from the input document and are used by a static IE model to identify the information. Expectations or self-criticism can help to identify highly confident information and relevant meta-features. Transfer knowledge is responsible for the projection or comparison of the given meta-features. The meta-features and transfer knowledge elements make up the dynamic IE model and are extended in an incremental process. The elements of the process model are adressed in more detail in the following:

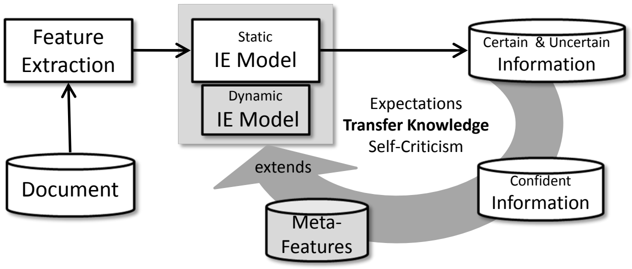


Fig. 3. Extended process model with meta-features and transfer knowledge.

Meta-Features Relations between features and information, respectively patterns, are explicitly implemented by meta-features. These are not only created for the extracted information, but also for possible candidates. A simple meta-feature, for example for the extraction of headlines, states that the bold feature indicates a headline in this document.

Expectations and Self-Criticism Since even only a single incorrect information can lead to a potentially high number of incorrect information, the correctness and confidence of an information is essential for the meta-level information extraction. There are two ways to identify an information suitable for the extraction of meta-features. If the knowledge engineer already has some assumptions about the content of the input documents, especially on the occurence of certain information, then these expectations can be formalized in order to increase the confidence of the information. In the absence of expecations, self-criticism of the IE model using features or a confidence value can highlight a suitable information. Furthermore, self-critism can be used to reduce the incorrect transfer of meta-features by rating newly identified information.

Transfer Knowledge The transfer knowledge models the human behaviour in practice and can be classified in three categories: *Agglomeration* knowledge processes multiple meta-features and creates new composed meta-features. Then, *projection* knowledge defines the transfer of the meta-features to possible candidates of new information. *Comparison* knowledge finally formalizes how the similarity of the meta-features of the original information and a candidate information is calculated. The usage of these different knowledge types in an actual process depends on the kind of repetitive structures and meta-features.

In section 3, specific examples of these elements are explained in the context of their application.

3 Case Studies

For a demonstation of the applicability and benefit of the approach, the two subtasks of the IE applications introduced earlier are addressed. The meta-level approach is realized with the rule-based TextMarker system and the statistical natural language processing toolkit ClearTK [5] is used for the the supervised machine learning methods CRF[1] and SVM[2]. The three methods operate in the same architecture (UIMA) and process the identical features. The same documents are applied for the training and test phase of the machine learning approaches and intentionally no k-fold cross evaluation is used, since it is hardly applicable for the knowledge engineering approach. Yet, the selected features do not amplify overfitting, e.g., no stem information is used. The evaluation of the SVM did not return reasonable values, probably because of the limited amount of documents and features in combination with the selected kernel method. Therefore, only results of the meta-level approach and CRF are presented using the F1-measure.

3.1 The TextMarker System

The TEXTMARKER system[3] is a rule-based tool for information extraction and text processing tasks [6]. It provides a full-featured development environment based on the DLTK framework[4] and a build process for UIMA Type Systems and generic UIMA

[1] The CRF implementation of Mallet (http://mallet.cs.umass.edu/) is used.

[2] The SVM implementation of SVMLight (http://svmlight.joachims.org/) is used.

[3] http://textmarker.sourceforge.net/

[4] http://www.eclipse.org/dltk/

Analysis Engines [1]. Different components for rule explanation and test-driven development [7] facilitate the knowledge engineering of rule-based information extraction components. The basic idea of the TEXTMARKER language is similar to JAPE [8]: rules match on combinations of predefined types of annotations and create new annotations. Furthermore, the TEXTMARKER language provides an extension mechanism for domain dependent language elements, several scripting functionalities and a dynamic view on the document. Due to the limited space, we refer to [6, 7] for a detailed description of the system.

3.2 CVs

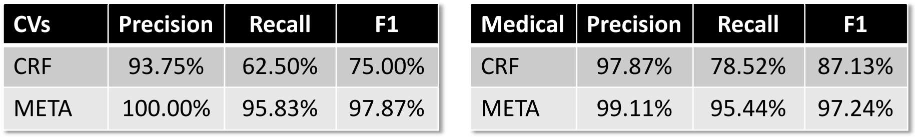
In this case study, we evaluate a subtask of the extraction of CV information: Companies in a past work experience of a person, respectively the employer. The corpus contains only 15 documents with 72 companies. The selected features consists of already extracted slots, layout information, simple token classes and a list of locations of one country. The meta-features are based on the position of confident information dependent on the layout and in relation to other slots. Agglomeration knowledge uses these meta-features to formalize a pattern of the common appearence of the companies. Then, projection knowledge uses this pattern to identify new information, that is rated by rules for self-criticism. In Figure 4, the results of the evaluation are listed. The meta-level approach achieved a F1-measure of 97.87% and the CRF method reached 75.00%. The low recall value of the CRF is caused by the limited amount of available features. However, the meta-level approach was able to compensate for this loss using the meta-features.

3.3 Medical discharge letters

A subtask of the extraction of information from medical discharge letters is the recognition of headlines. In oder to evaluate the approaches we use a corpus with 141 documents and 1515 headlines. The extracted features consist mainly of simple token classes and layout information, e.g., bold, underlined, italic and freeline. In this case study, the expectation to find a *Diagnose* or *Anamnese* headline is used to identify a confident information. Then, meta-features describing its actual layout are created and transferred by projection knowledge. Finally, comparison knowledge is used to calculate the similarity of the layout of the confident information and a candidate for a headline. The results of the evaluation are shown in figure 4: The meta-level approach was evaluated with 97.24% and the CRF method achieved a F1-measure of 87.13%. CRF extracted the same headlines as the meta-level approach in many documents. However, the conflicting layout styles of the some authors caused, as expected, a high number of false negative errors resulting in a lower recall value.

3.4 Related Work and Discussion

In the case studies, we have seen that the proposed approach performs very promissing and achieves considerably better accuracy measures than the standard approach using



CVs	Precision	Recall	F1
CRF	93.75%	62.50%	75.00%
META	100.00%	95.83%	97.87%

Medical	Precision	Recall	F1
CRF	97.87%	78.52%	87.13%
META	99.11%	95.44%	97.24%

Fig. 4. Results of the CVs and medical discharge letters evaluation

machine learning techniques, that is CRF. The machine learning methods would potentially perform better using more (and/or 'better') features, however, the same is true for the meta-level IE approach. The approach is not only very effective but also rather efficient, since the proposed approach required only about 1-2 hours for formalizing the necessary meta-features and transfer knowledge, significantly less time than the time spent for the annotation of the examples.

To the best of the authors' knowledge, the approach is novel in the IE community and application. However, similar ideas to the core idea of transfering features have been adressed in the feature construction and inductive logic programming community, e.g., [9]. However, in this context there is no direct 'feedback' according to a certain process, and also no distinction between meta-features and transfer knowledge that is provided by the presented approach. According to the analogy of human reasoning, it is often easier to formalize each knowledge element separately. Especially in information extraction, there are approaches using extracted information in a meta-learning process, e.g., [10]. However, compared to our approach no meta-features dependent on extracted information and no transfer knowledge is used. The proposed approach is able to adapt to peculiarities of certain authors of the documents, similarly to the adaptation phase of common speech processing and speech understanding systems.

4 Conclusions

In this paper, we have presented a meta-level information extraction approach that extends the common IE process model by including meta-level features. These meta-features are created using already extracted information, e.g., given by repetitive structural constructs of the present feature space. We have described a general model for the application of the presented approach, and we have demonstrated its benefit in two case studies utilizing a rule-based system for information extraction.

For future work, the exchange of transfer knowledge and meta-features between documents can further enrich the process model in specific domains. We plan to extend the approach in order to incorporate the automatic acquisition of transfer knowledge. Techniques from inductive logic programming [11] can potentially provide helpful methods and support the knowledge engineer to automatically acquire the needed transfer and meta knowledge. For the automatic acquisition, self criticism capabilities and the inclusion of the expecations of the developer are essential. Finally, we are also planning to combine the approach with machine learning methods, like SVM and CRF [5]:

Acknowledgements

This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-2.

References

1. Ferrucci, D., Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Nat. Lang. Eng. **10**(3-4) (2004) 327–348
2. McCallum, A., Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In: Proc. of the seventh conference on Natural language learning at HLT-NAACL 2003, Morristown, NJ, USA, Association for Computational Linguistics (2003) 188–191
3. Turmo, J., Ageno, A., Català, N.: Adaptive Information Extraction. ACM Comput. Surv. **38**(2) (2006) 4
4. Li, D., Savova, G., Kipper-Schuler, K.: Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. In: Proc. of the Workshop on Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, Association for Computational Linguistics (June 2008) 94–95
5. Ogren, P.V., Wetzler, P.G., Bethard, S.: ClearTK: A UIMA Toolkit for Statistical Natural Language Processing. In: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC). (2008)
6. Atzmueller, M., Kluegl, P., Puppe, F.: Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. In: Proc. of the LWA-2008 (KDML Track). (2008) 1–7
7. Kluegl, P., Atzmueller, M., Puppe, F.: Test-Driven Development of Complex Information Extraction Systems using TextMarker. In: KESE at KI 2008. (2008)
8. Cunningham, H., Maynard, D., Tablan, V.: JAPE: A Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield (November 2000)
9. Flach, P.A., Lavrac, N.: The Role of Feature Construction in Inductive Rule Learning. In Raedt, L.D., Kramer, S., eds.: Proc. ICML2000 Workshop on Attribute-Value and Relational Learning: crossing the boundaries, 17th International Conference on Machine Learning (July 2000) 1–11
10. Sigletos, G., Paliouras, G., Spyropoulos, C.D., Stamatopoulos, T.: Meta-Learning beyond Classification: A Framework for Information Extraction from the Web. In: Proc. of the Workshop on Adaptive Text Extraction and Mining. The 14th Euro. Conf. on Machine Learning and the 7th Euro. Conf. on Principles and Practce of knowledge Discovery in Databases. (2003)
11. Thomas, B.: Machine Learning of Information Extraction Procedures - An ILP Approach. PhD thesis, Universität Koblenz-Landau (2005)