Constraint-driven Evaluation in UIMA Ruta

Andreas Wittek[1], Martin Toepfer[1], Georg Fette[12],
Peter Kluegl[12], and Frank Puppe[1]

[1] Department of Computer Science VI, University of Wuerzburg,
Am Hubland, Wuerzburg, Germany
[2] Comprehensive Heart Failure Center, University of Wuerzburg,
Straubmuehlweg 2a, Wuerzburg, Germany
{a.wittek,toepfer,fette,pkluegl,puppe}@informatik.uni-wuerzburg.de

Abstract. This paper presents an extension of the UIMA Ruta Workbench for estimating the quality of arbitrary information extraction models on unseen documents. The user can specify expectations on the domain in the form of constraints, which are applied in order to predict the F_1 score or the ranking. The applicability of the tool is illustrated in a case study for the segmentation of references, which also examines the robustness for different models and documents.

1 Introduction

Apache UIMA [5] and the surrounding ecosystem provide a powerful framework for engineering state-of-the-art Information Extraction (IE) systems, e.g., in the medical domain [13]. Two main approaches for building IE models can be distinguished. One approach is based on manually defining a set of rules, e.g., with UIMA Ruta[3] (Rule-based Text Annotation) [7][4], that is able to identify the interesting information or annotations of specific types. A knowledge engineer writes, extends, refines and tests the rules on a set of representative documents. The other approach relies on machine learning algorithms, such as probabilistic graphical models like Conditional Random Fields (CRF) [10]. Here, a set of annotated gold documents is used as a training set in order to estimate the parameters of the model. The resulting IE system of both approaches, the statistical model and the set of rules, is evaluated on an additional set of annotated documents in order to estimate its accuracy or F_1 score, which is then assumed to hold for the application in general. However, while the system performed well in the evaluation setting, its accuracy decreases when applied on unseen documents, maybe because the set of documents applied for developing the IE system was not large or not representative enough. In order to estimate the actual performance, either more data is labeled or the results are manually checked by a human, who is able to validate the correctness of the annotations.

Annotated documents are essential for developing IE systems, but there is a natural lack of labeled data in most application domains and its creation is

[3] http://uima.apache.org/ruta.html
[4] previously published as TextMarker

error-prone, cumbersome and time-consuming as is the manual validation. An automatic estimation of the IE system's quality on unseen documents would therefore provide many advantages. A human is able to validate the created annotations using background knowledge and expectations on the domain. This kind of knowledge is already used by current research in order to improve the IE models (c.f. [1, 6, 11]), but barely to estimate IE system's quality.

This paper introduces an extension of the UIMA Ruta Workbench for exactly this use case: Estimating the quality and performance of arbitrary IE models on unseen documents. The user can specify expectations on the domain in the form of constraints thus the name Constraint-driven Evaluation (CDE). The constraints rate specific aspects of the labeled documents and are aggregated to a single CDE score, which provides a simple approximation of the evaluation measure, e.g., the token-based F_1 score. The framework currently supports two different kinds of constraints: Simple UIMA Ruta rules, which express specific expectations concerning the relationship of annotations, and annotation-distribution constraints, which rate the coverage of features. We distinguish two tasks: predicting the actual F_1 score of a document and estimating the ranking of the documents specified by the actual F_1 score. The former task can give answers on how good the model performs. The latter task points to documents where the IE model can be improved. We evaluate the proposed tool in a case study for the segmentation of scientific references, which tries to estimate the F_1 score of a rule-based system. The expectations are additionally applied on documents of a different distribution and on documents labeled by a different IE model. The results emphasize the advantages and usability of the approach, which works already with minimal efforts due to a simple fact: It is much easier to estimate how good a document is annotated than to actually identify the positions of defective or missing annotations.

The rest of the paper is structured as follows. In the upcoming section, we describe how our work relates to other fields of Information Extraction research. We explain the proposed CDE approach in Section 3. Section 4 covers the case study and the corresponding results. We conclude with pointers to future work in Section 5.

2 Related Work

Besides standard classification methods, which fit all model parameters against the labeled data of the supervised setting, there have been several efforts to incorporate background knowledge from either user expectations or external data analysis. Bellare et al. [1], Graça et al. [6] and Mann and McCallum [11], for example, showed how moments of auxiliary expectation functions on unlabeled data can be used for such a purpose with special objective functions and an alternating optimization procedure. Our work on constraint-driven evaluation is partly inspired by this idea, however, we address a different problem. We suggest to use auxiliary expectations to estimate the quality of classifiers on unseen data.

A classifier's confidence describes the degree to which it believes that its own decisions are correct. Several classifiers provide intrinsic measures of confidence, for example, naive Bayes classifiers. Culotta and McCallum [4], for instance, studied confidence estimation for information extraction. They focus on predictions about field and record correctness of single instances. Their main motivation is to filter high precision results for database population. Similar to CDE, they use background knowledge features like record length, single field label assignments and field confidence values to estimate record confidence. CDE generalizes common confidence estimation because the goal of CDE is the estimation of the quality of arbitrary models.

Active learning algorithms are able to choose the order in which training examples are presented in order to improve learning, typically by selective sampling [2]. While the general CDE setting does not necessarily contain aspects of selective sampling, consider for example the batch F_1 score prediction task, the ranking task can be used as a selective sampling strategy in applications to find instances that support system refactoring. The focus of the F_1 ranking task, however, still differs from active learning goals which is essential for the design of such systems. Both approaches are supposed to favor different techniques to fit their different objectives. Popular active learning approaches such as density-weighting (e.g., [12]) focus on dense regions of the input distribution. CDE, however, tries to estimate the quality of the model on the whole data set and hence demands for differently designed methods. Despite their differences, the combination of active learning and CDE would be an interesting subject for future work. CDE may be used to find weak learners of ensembles and informative instances for these learners.

3 Constraint-driven Evaluation

The Constraint-driven Evaluation (CDE) framework presented in this work allows the user to specify expectations about the domain in form of constraints. These constraints are applied on documents with annotations, which have been created by an information extraction model. The results of the constraints are aggregated to a single CDE score, which reflects how well the annotations fulfill the user's expectations and thus provide a predicted measurement of the model's quality for these documents. The framework is implemented as an extension of the UIMA Ruta Workbench. Figure 1 provides a screenshot of the CDE perspective, which includes different views to formalize the set of constraints and to present the predicted quality of the model for the specified documents.

We define a constraint in this work as a function $C : \text{CAS} \rightarrow [0, 1]$, which returns a confidence value for an annotated document (CAS) where high values indicates that the expectations are fulfilled. Two different types of constraints are currently supported: RULE constraints are simple UIMA Ruta rules without actions and allow to specify sequential patterns or other relationships between annotations that need to be fulfilled. The result is basically the ratio of how often the rule has tried to match compared to how often the rule has actually

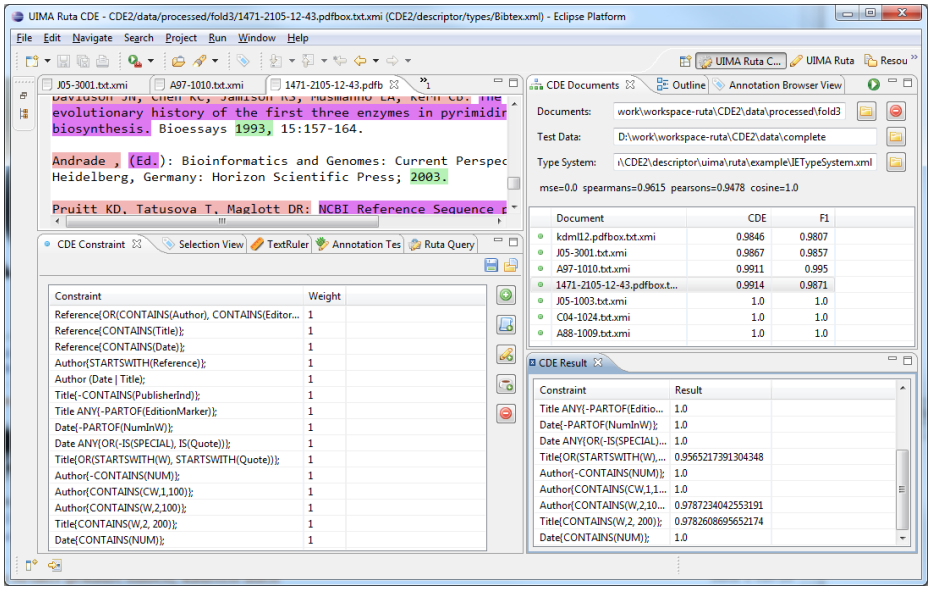

Fig. 1. CDE perspective in the UIMA Ruta Workbench. Bottom left: Expectations on the domain formalized as constraints. Top right: Set of documents and their CDE scores. Bottom right: Results of the constraints for the selected document.

matched. An example for such a constraint is `Document{CONTAINS(Author)};`, which specifies that each document must contain an annotation of the type *Author*. The second type of supported constraints are Annotation Distribution (AD) constraints (c.f. Generalized Expectations [11]). Here, the expected distribution of an annotation or word is given for the evaluated types. The result of the constraint is the cosine similarity of the expected and the observed presence of the annotation or word within annotations of the given types. A constraint like `"Peter": Author 0.9, Title 0.1`, for example, indicates that the word "Peter" should rather be covered by an *Author* annotation than by a *Title* annotation. The set of constraints and their weights can be defined using the *CDE Constraint* view (c.f. Figure 1, bottom left).

For a given set of constraints $C = \{C_1, C_2 ... C_n\}$ and corresponding weights $w = \{w_1, w_2, ..., w_n\}$, the CDE score for each document is defined by the weighted average:

$$\text{CDE} = \frac{1}{n} \sum_{i}^{n} w_i \cdot C_i \tag{1}$$

The CDE scores for a set of documents may already be very useful as a report how well the annotations comply with the expectations on the domain. However, one can further distinguish two tasks for CDE: the prediction of the actual evaluation score of the model, e.g., the token-based F_1 score, and the

prediction of the quality ranking of the documents. While the former task can give answers how good the model performs or whether the model is already good enough for the application, the latter task provides a useful tool for introspection: Which documents are poorly labeled by the model? Where should the model be improved? Are the expectations on the domain realistic? Due to the limited expressiveness of the aggregation function, we concentrate on the latter task. The CDE scores for the annotated documents are depicted in the *CDE Documents* view (c.f. Figure 1, top right). The result of each constraint for the currently selected document is given in the *CDE Results* view (c.f. Figure 1, bottom right).

The development of the constraints needs to be supported by tooling in order to achieve an improved prediction in the intended task. If the user extends or refines the expectations on the domain, then a feedback whether the prediction has improved or deteriorated is very valuable. For this purpose, the framework provides functionality to evaluate the prediction quality of the constraints itself. Given a set of documents with gold annotations, the CDE score of each document can be compared to the actual F_1 score. Four measures are applied to evaluate the prediction quality of the constraints: the mean squared error, the Spearman's rank correlation coefficient, the Pearson correlation coefficient and the cosine similarity. For optimizing the constraints to approximate the actual F_1 score, the Pearson's r is maximized, and for improving the predicted ranking, the Spearman's ρ is maximized. If documents with gold annotations are available, then the F_1 scores and the values of the four evaluation measures are given in the *CDE Documents* view (c.f. Figure 1, top right).

4 Case Study

The usability and advantages of the presented work are illustrated with a simple case study concerning the segmentation of scientific references, a popular domain for evaluating novel information extraction models. In this task, the information extraction model normally identifies about 12 different entities of the reference string, but in this case study we limited the relevant entities to *Author*, *Title* and *Date*, which are commonly applied in order to identify the cited publication.

In the main scenario of the case study, we try to estimate the extraction quality of a set of UIMA Ruta rules that shall identify the *Author*, *Title* and *Date* of a reference string. For this purpose, we define constraints representing the background knowledge about the domain for this specific set of rules. Additionally to this main setting of the case study, we also measure the prediction of the constraints in two different scenarios: In the first one, the documents have been labeled not by UIMA Ruta rules, but by a CRF model [10]. The CRF model was trained with a limited amount of iterations in a 5-fold manner. In a second scenario, we apply the UIMA Ruta rules on a set of documents of a different distribution including unknown style guides.

Table 1 provides an overview of the applied datasets. We make use of the references dataset of [9]. This data set is homogeneously divided in three sub-datasets with respect to their style guides and amount of references, which are

Table 1. Overview of utilized datasets.

D_{ruta}	219 references in 8 documents used to develop the set of UIMA Ruta rules.
D_{dev}	192 references in 8 documents labeled by the UIMA Ruta rules and applied for developing the constraints.
D_{test}	155 references in 7 documents labeled by the UIMA Ruta rules and applied to evaluate the constraints.
D_{crf}	D_{ruta}, D_{dev} and D_{test} (566 references in 23 documents) labeled by a (5-fold) CRF model.
D_{gen}	452 references in 28 documents from a different source with unknown style guides labeled by the UIMA Ruta rules.

applied to develop the UIMA Ruta rules, define the set of constraints, and to evaluate the prediction of the constraints compared to the actual F_1 score. The CRF model is trained on the partitions given in [9]. The last dataset D_{gen} consists of a mixture of the datasets CORA, CITESEERX and FLUX-CIM described in [3] generated by the rearrangement of [8].

Table 2. Overview of evaluated sets of constraints.

C_{ruta}	15 RULE constraints describing general expectations for the entities *Author*, *Title* and *Date*. The weight of each constraint is set to 1.
$C_{ruta+bib}$	C_{ruta} extended with one additional AD constraint covering the entity-distribution of words extracted from Bibsonomy. The weight of each constraint is set to 1.
$C_{ruta+5xbib}$	Same set of constraints as in $C_{ruta+bib}$, but the weight of the additional AD constraint is set to 5.

Table 2 provides an overview of the different sets of constraints, whose predictions are compared to the actual F_1 score. First, we extended and refined a set of UIMA Ruta rules until they achieved an F_1 score of 1.0 on the dataset D_{ruta}. Then, 15 RULE constraints C_{ruta}[5] have been specified using the dataset D_{dev}. The definition of the UIMA Ruta rules took about two hours and the definition of the constraints about one hour. Additionally to the RULE constraints, we created an AD constraint, which consists of the entity distribution of words that occurred at least 1000 times in the latest Bibtex database dump of Bibsonomy[6]. The set of constraints $C_{ruta+bib}$ and $C_{ruta+5xbib}$ combine both types of constraints with different weighting.

Table 3 contains the evaluation, which compares the predicted CDE score to the actual token-based F_1 score for each document. We apply two different

[5] The actual implementation of the constraints as UIMA Ruta rules is depicted in Figure 1 (lower left part).

[6] `http://www.kde.cs.uni-kassel.de/bibsonomy/dumps`

Table 3. Spearman's ρ and Pearson's r given for the predicted CDE score (for each document) compared to the actual F_1 score.

	C_{ruta}		$C_{ruta+bib}$		$C_{ruta+5xbib}$	
Dataset	ρ	r	ρ	r	ρ	r
D_{dev}	0.8708	0.9306	0.9271	0.9405	0.8051	0.6646
D_{test}	0.9615	0.9478	0.9266	0.8754	0.8154	0.6758
D_{crf}	0.6793	0.7881	0.7429	0.8011	0.7117	0.7617
D_{gen}	0.7089	0.8002	0.7724	0.8811	0.8150	0.9504

correlation coefficients for measuring the quality of the prediction: Spearman's ρ gives an indication about the ranking of the documents and Pearson's r provides a general measure of linear dependency.

Although the expectations defined by the sets of constraints are limited and quite minimalistic covering mostly only common expectations, the results indicate that they can be useful in any scenario. The results for dataset D_{dev} are only given for completeness since this dataset was applied to define the set of constraints. The results for the dataset D_{test}, however, reflect the prediction on unseen documents of the same distribution. The ranking of the documents was almost perfectly estimated with a Spearman's ρ of 0.9615[7]. The coefficients for the other scenarios D_{crf} and D_{gen} are considerably decreased, but the CDE scores are nevertheless very useful for an assessment of the extraction model's quality. The five worst documents in D_{gen} (including new style guides), for example, have been reliably detected. The results show that the AD constraints can improve the prediction, but do not exploit their full potential in the current implementation. The impact measured for the dataset D_{crf} is not as distinctive since the CRF model already includes such features and thus is able to avoid errors that are detected by these constraints. However, the prediction in the dataset D_{gen} is considerably improved. The UIMA Ruta rules produce severe errors in documents with new style guides, which are easily detected by the word distribution.

5 Conclusions

This paper presented a tool for the UIMA community implemented in UIMA Ruta, which enables to estimate the extraction quality of arbitrary models on unseen documents. Its introspective report is able to improve the development of information extraction models already with minimal efforts. This is achieved by formalizing the background knowledge about the domain with different types of constraints. We have shown the usability and advantages of the approach in a case study about segmentation of references. Concerning future work, many prospects for improvements remain, for example a logistic regression model for

[7] The actual CDE and F_1 scores of D_{test} are depicted in Figure 1 (right part)

approximating the scores of arbitrary evaluation measures, new types of constraints, or approaches to automatically acquire the expectations on a domain.

Acknowledgments This work was supported by the Competence Network Heart Failure, funded by the German Federal Ministry of Education and Research (BMBF01 EO1004).

References

1. Bellare, K., Druck, G., McCallum, A.: Alternating Projections for Learning with Expectation Constraints. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in AI. pp. 43–50. AUAI Press (2009)
2. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning 15, 201–221 (1994)
3. Councill, I., Giles, C.L., Kan, M.Y.: ParsCit: an Open-source CRF Reference String Parsing Package. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). ELRA, Marrakech, Morocco (2008)
4. Culotta, A., McCallum, A.: Confidence Estimation for Information Extraction. In: Proceedings of HLT-NAACL 2004: Short Papers. pp. 109–112. HLT-NAACL-Short '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
5. Ferrucci, D., Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering 10(3/4), 327–348 (2004)
6. Graca, J., Ganchev, K., Taskar, B.: Expectation Maximization and Posterior Constraints. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) NIPS 20, pp. 569–576. MIT Press, Cambridge, MA (2008)
7. Kluegl, P., Atzmueller, M., Puppe, F.: TextMarker: A Tool for Rule-Based Information Extraction. In: Chiarcos, C., de Castilho, R.E., Stede, M. (eds.) Proceedings of the 2nd UIMA@GSCL Workshop. pp. 233–240. Gunter Narr Verlag (2009)
8. Kluegl, P., Hotho, A., Puppe, F.: Local Adaptive Extraction of References. In: 33rd Annual German Conference on Artificial Intelligence (KI 2010). Springer (2010)
9. Kluegl, P., Toepfer, M., Lemmerich, F., Hotho, A., Puppe, F.: Collective Information Extraction with Context-Specific Consistencies. In: Flach, P.A., Bie, T.D., Cristianini, N. (eds.) ECML/PKDD (1). Lecture Notes in Computer Science, vol. 7523, pp. 728–743. Springer (2012)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning pp. 282–289 (2001)
11. Mann, G.S., McCallum, A.: Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. J. Mach. Learn. Res. 11, 955–984 (2010)
12. McCallum, A., Nigam, K.: Employing EM and Pool-Based Active Learning for Text Classification. In: Shavlik, J.W. (ed.) ICML. pp. 350–358. Morgan Kaufmann (1998)
13. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA 17(5), 507–513 (Sep 2010)