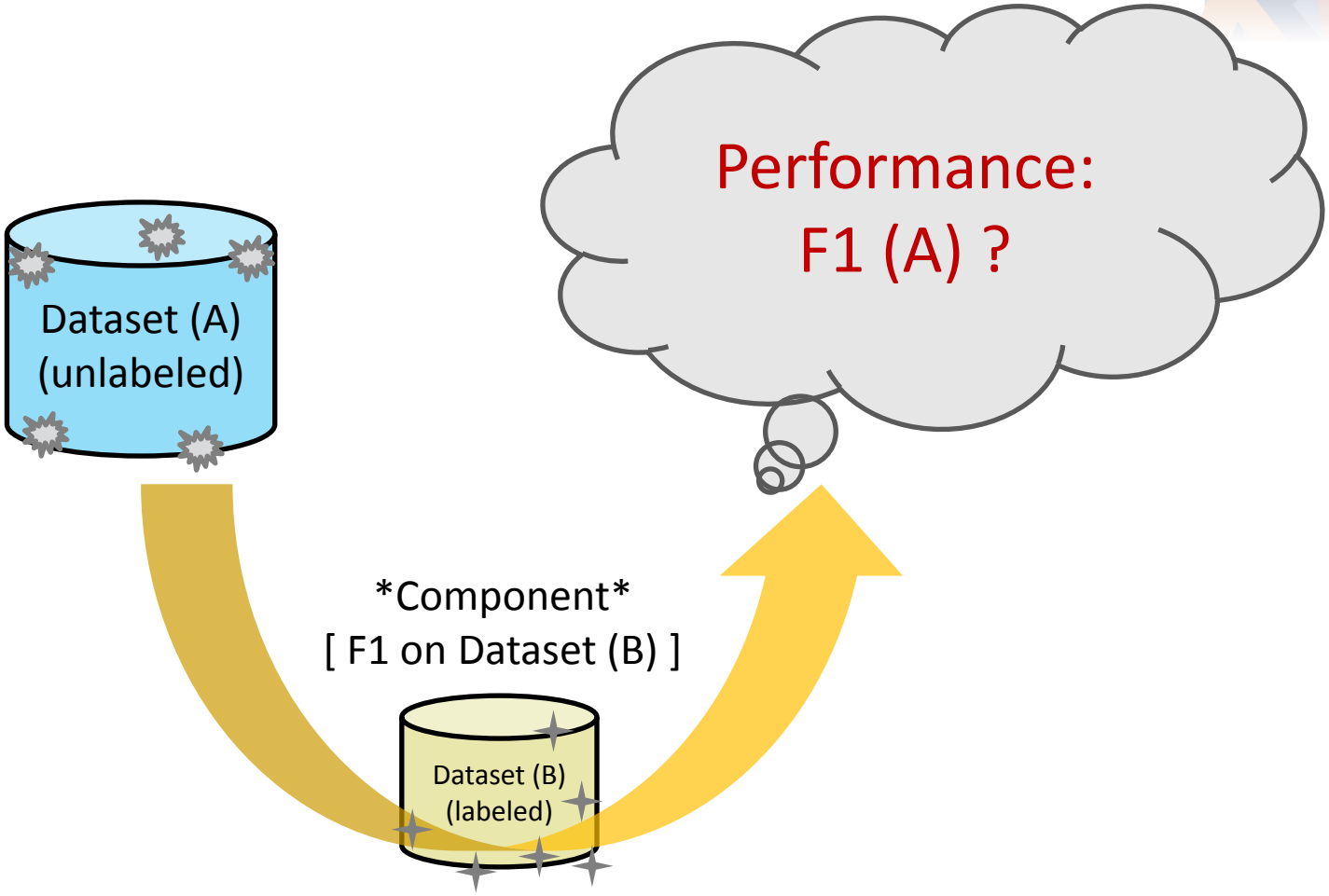
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Constraint-driven Evaluation in UIMA Ruta

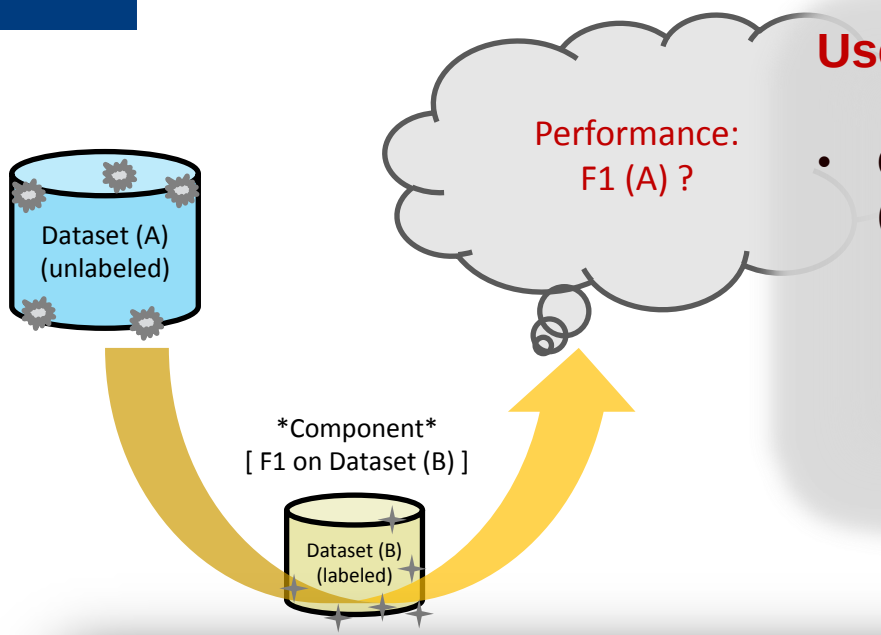
Estimating the quality of arbitrary models on unseen documents.

Andreas Wittek, Martin Toepfer, Georg Fette, Peter Klügl, Frank Puppe

Constraint-driven Evaluation



Constraint-driven Evaluation



Use Expectations:

- Case Study: Segmentation of References (Extracting BibTeX fields)
  - References should have a title containing more than 2 words
  - References should have 1! author field
  - „VHDL“ should be labeled as „title“

R. Lipsett, C. Schaefer, C. Ussery, VHDL: Hardware Description and Design, Kluwer Academic Publishers, 1989.
T. Imielinski, S. Viswanathan, Adaptive Wire less Information Systems, Proc. of SIGDBS Conf., Tokyo, October, 1994.
K. Chandy and J. Misra, Parallel Program Design: A Foundation, Addison-Wesley, 1988.

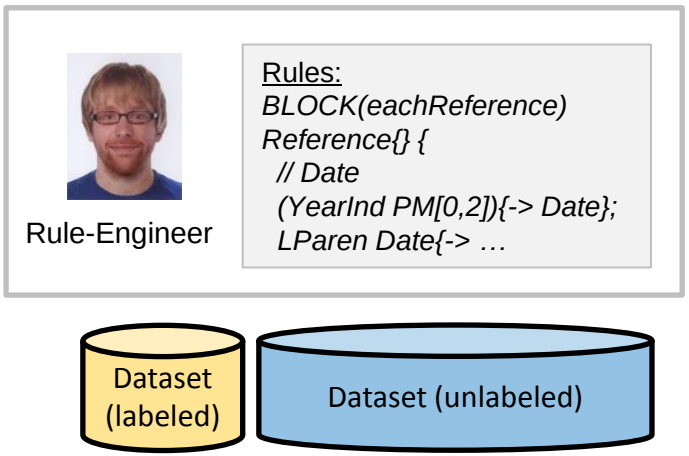
title , author

Good reasons for Constrained-driven Evaluation (CDE)

Choose one of several 3rd-party components, Debug rule-based systems, Decision support for make-or-buy

Reason 1: Development Support

- typical scenario in NLP applications:



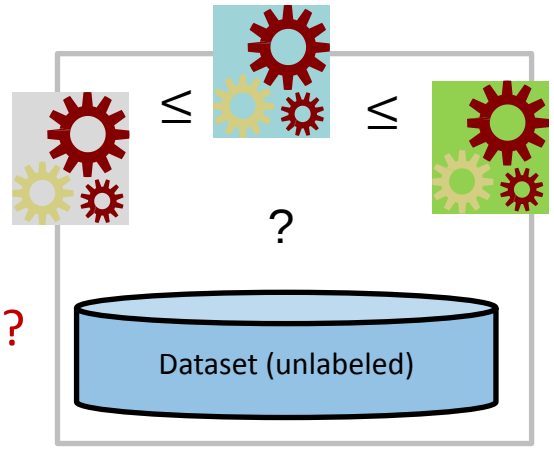
How good is my system at the moment?

Which documents should I look at to refine my system?

- **Constrained-driven evaluation (CDE) provides:**
  - **Ranking of unlabeled documents according to the expected quality of a component**
  - **F1 score prediction on unlabeled data sets**

Reason 2: Component Selection

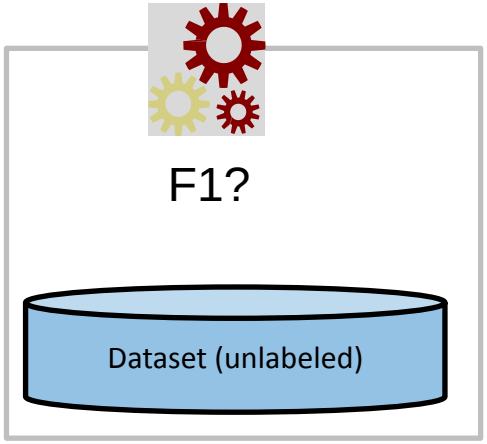
- typical scenario in NLP applications:
  - several 3rd-party components
    (e.g., CRF & SVM & Rule-based NER model trained on CoNLL)
  - Unlabeled data-set
    (e.g., a custom newspaper corpus)
  - Which component is the best (relatively)?



- **Constrained-driven evaluation (CDE):**
  - **Ranking of components according to performance estimation of arbitrary models using constraints**

Reason 3: make-or-buy

- typical scenario in NLP applications:
  - 3rd-party component
    (e.g., CRF NER model trained on CoNLL)
  - Unlabeled data-set
    (e.g., a custom newspaper corpus)



  - Are the results good enough for my application? Or do I have to develop my own system?
    (absolute quality estimation)
- **Constrained-driven evaluation (CDE):**
  - **Automated performance estimation of arbitrary models with formalized expectations**

CDE Tasks

- **F1 Prediction** task: Estimate F1-score of arbitrary models on unlabeled data
  - Reason 2 (component selection): *relative* performance
  - Reason 3 (make-or-buy) *absolute* performance
  - [Reason 1 (development support)]
- **Ranking** task: Rank documents wrt. F1
  - Reason 1 (development support)

CDE Workflow

- *Basic* operations for rapid component development:
  1. Gather constraints
     - **Rule Constraints**:
       formalize background-knowledge of domain experts
     - **Annotation distribution (AD) Constraints**:
       Collect statistics (using large data-bases)
  2. Constraint-driven Evaluation
     - Apply model on unlabeled data-set
     - Apply constraints on unlabeled data-set
     - Compute aggregate **CDE score** for each document

CDE Score

- reflects how well the output of the component complies with the constraints:

$$\mathrm{CDE} = \frac{1}{\alpha}\sum_{i=1}^{n} w_i C_i$$

where w_i is a weight, α is a normalizing constant, and $C_i \in [0,1]$ is either

  - a *rule constraint score*:
    #rule matched / #rule has tried to match
  - or an *annotation distribution (AD) constraint score*:
    cosine similarity between expected and observed frequency

Rule Constraints

- Formalize expectations of domain experts as UIMA Ruta rules
- Example:

```
Arthur M. Keller and Julie Basu. A predicate-based caching scheme for
client-server database architectures. In Proceedings of PDIS-94, 1994.

Buntine, W. (1994). Operations for learning with graphical models.
Journal of Artificial Intelligence Research, 2, 159-225.
```

Author (Title | Year);

Author {-CONTAINS(NUM)};

Title {CONTAINS(W,2,200)};

Rule Constraints

- *Title {CONTAINS(W,2,200)};*
  → score: 4/10 (4 out of 10 titles have ≥2 words)

Constraint satisfied?

n
n
n
y
y
y
n
n
n
y

A. Said and W.A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," IEEE Transactions on Ciruits and Systems for Video Technology, vol. 6, no. 3, pp. 243--249, June 1996.
M. Yang and L. M. Ni, "Incremental design of scalable interconnection networks using basic building blocks," Symposium of Parallel and Distributed Processing, pp. 252 -- 259, Oct. 1995.
L. Hernquist, "Vectorization of Tree Traversals," Journal of Computational Physics, vol. 87, 1990.
J. Hipp, R. Lober, "Plastering: All-Hexahedral Mesh Generation Through Connectivity Resolution," Proc. 3 rd Int. Meshing Roundtable, 1994.
P. Sarkar, "An Iterative Algorithm for Optimal Style-Conscious Field Classification," Proc. 16th Int'l Conf. Pattern Recognition, pp. 243-246, Aug. 2002.
C. Rosene, "Incremental dependence analysis," Ph.D. thesis, Rice University, March 1990.
H. J. Landau and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis, and uncertainty—III: The dimension of the space of essentially time- and band-limited signals," Bell Syst. Tech. J., vol. 41, no. 4, pp. 1295–1336, Jul. 1962.

J. Carter and M. Wegman, "Universal classes of hash functions," J. of Computer and System Sciences, vol. 18, no. 2, pp. 143-154, 1979.
D. E. Whitney, "Force feedback control of manipulator fine motions," ASME J. Dyn. Syst. Meas. Contr., vol. 99, no. 2, pp. 91–97, 1977.
C. Norris and L. Pollock, "A scheduler-sensitive global register allocator," In Proc. of Supercomputing, pages 804-813, 1993.
M. J. Harrold and M. L. Soffa, "An incremental approach to unit testing during maintenance," Proceedings of the International Conference on Software Maintenance, pp. 362-367, October 1988.
J. H. M. J. Daily and K. Reiser, "Detecting obstacles in range imagery," Proc. of [ARPA] Image Understanding Workshop, pp. 87-97, 1987.
T. Klemenschits, & E. Bonek, "Radio Coverage of Road Tunnels at 900 and 1800 MHz

Some errors are detected correctly by this Rule Constraint;
Some errors are not detected

AD constraints may help in such cases (see next slide)

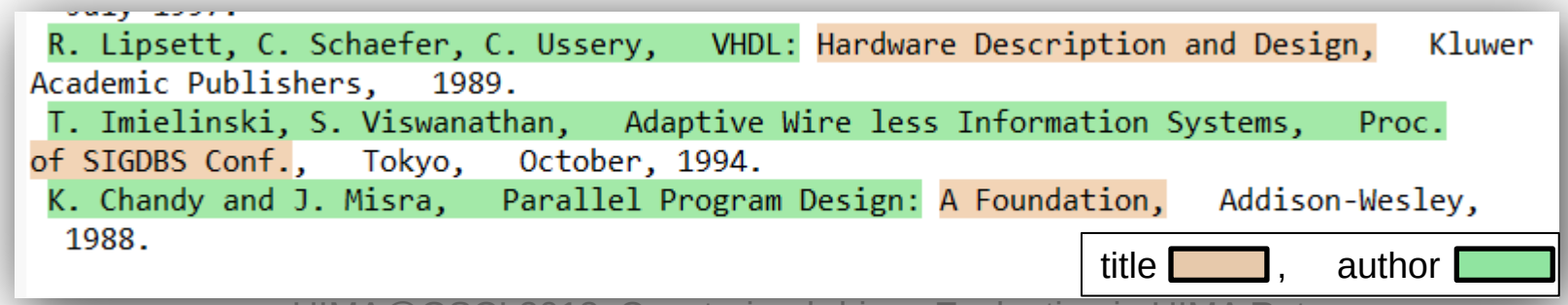
title

Annotation Distribution Constraints

- Statistical assumptions gathered from databases, e.g., Bibsonomy dump
- Example: in DB:
  “VHDL": Author 0.001, Title 0.4, …

  Word Label Freq.
- in Document:
  “VHDL"/Author: 1/1

$$C_i = \cos\left(\begin{pmatrix}1.0\\ \vdots\end{pmatrix}, \begin{pmatrix}0.001\\ \vdots\end{pmatrix}\right) = 0.51$$



Constrained-driven Evaluation

- **Extended** workflow:
  - Gather constraints
    - …
    - Constraint Development
    - Create a small dev.-set
    - Test constraints; specify weights
  - Constraint-driven Evaluation
    - …

Constraint Development

- measures to compare CDE score against ground-truth F1-score/ranking on a set of labeled documents:
  - Spearman (ranking)
  - Pearson (linear dependency)

Experimental Study

Segmentation of References of Scientific Papers

Setting

- D_{ruta} : used to develop rules for reference segmentation (219 references in 8 docs)
- D_{dev} : labeled by the constructed rules; used for developing the constraints (192 references in 8 docs)
- D_{test} : labeled by the constructed rules; used to evaluate the constraints (155 references in 7 documents)
- D_{crf} : labeled by training and applying CRFs (5-fold-cross) (ruta+dev+test: 566 references)
- D_{gen} : different source; unknown style guides; labeled by the constructed rules (452 references in 28 documents)

Setting

- Three constraint sets:
  1. C_{ruta}:
     15 Rule constraints (Author, Title, Date); weight of each constraint is 1
  2. $C_{\text{ruta+bib}}$:
     C_{ruta} + AD constraints (entity distribution of words extracted from Bibsonomy); weight of each constraint is set to 1
  3. $C_{\text{ruta+5*bib}}$:
     weight of each AD constraint set to 5

Results

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.

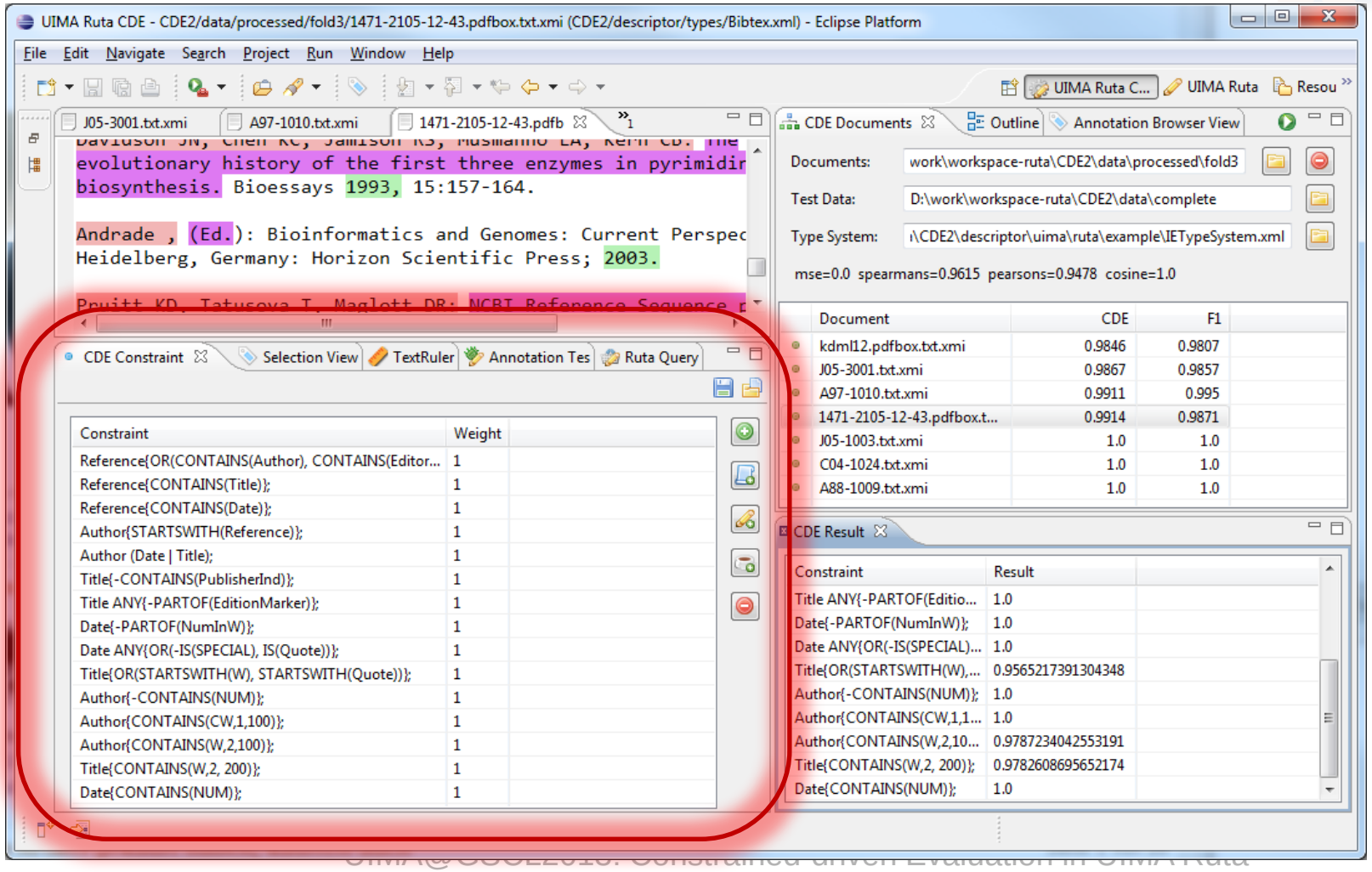
	C_{ruta}		$C_{ruta+bib}$		$C_{ruta+5xbib}$	
Dataset	ρ	r	ρ	r	ρ	r
D_{dev}	0.8708	0.9306	0.9271	0.9405	0.8051	0.6646
D_{test}	0.9615	0.9478	0.9266	0.8754	0.8154	0.6758
D_{crf}	0.6793	0.7881	0.7429	0.8011	0.7117	0.7617
D_{gen}	0.7089	0.8002	0.7724	0.8811	0.8150	0.9504

- Strong correlation ($\rho \geq 0.6, r \geq 0.6$ for all D_i, C_j):
  - constraint rules really estimate the performance of the models on the new unlabeled data
- Even with different data (new styles):
  - use AD constraints in this case

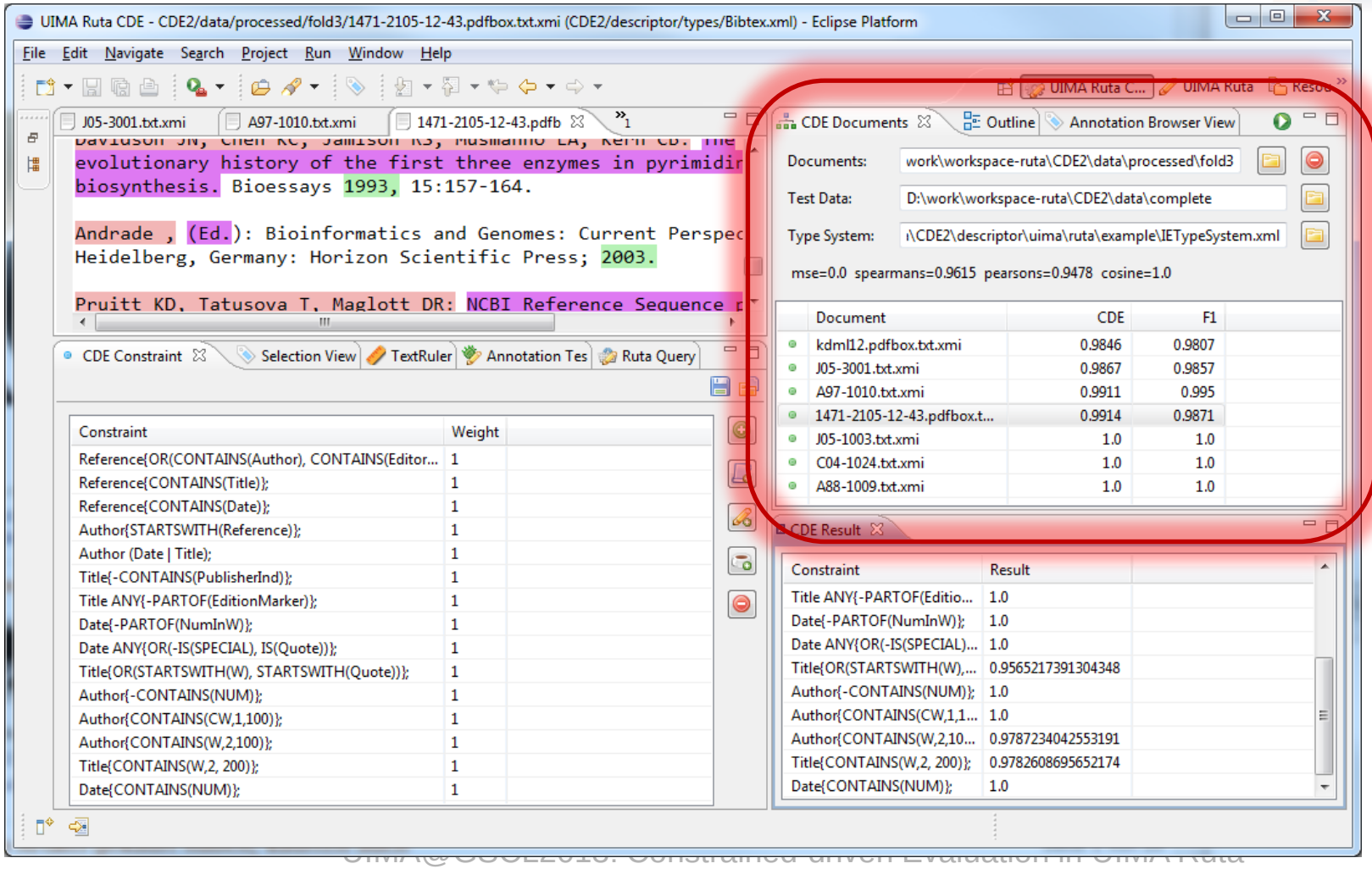
UIMA Ruta CDE Plugin

Constrained-driven Evaluation in the UIMA Ruta Workbench

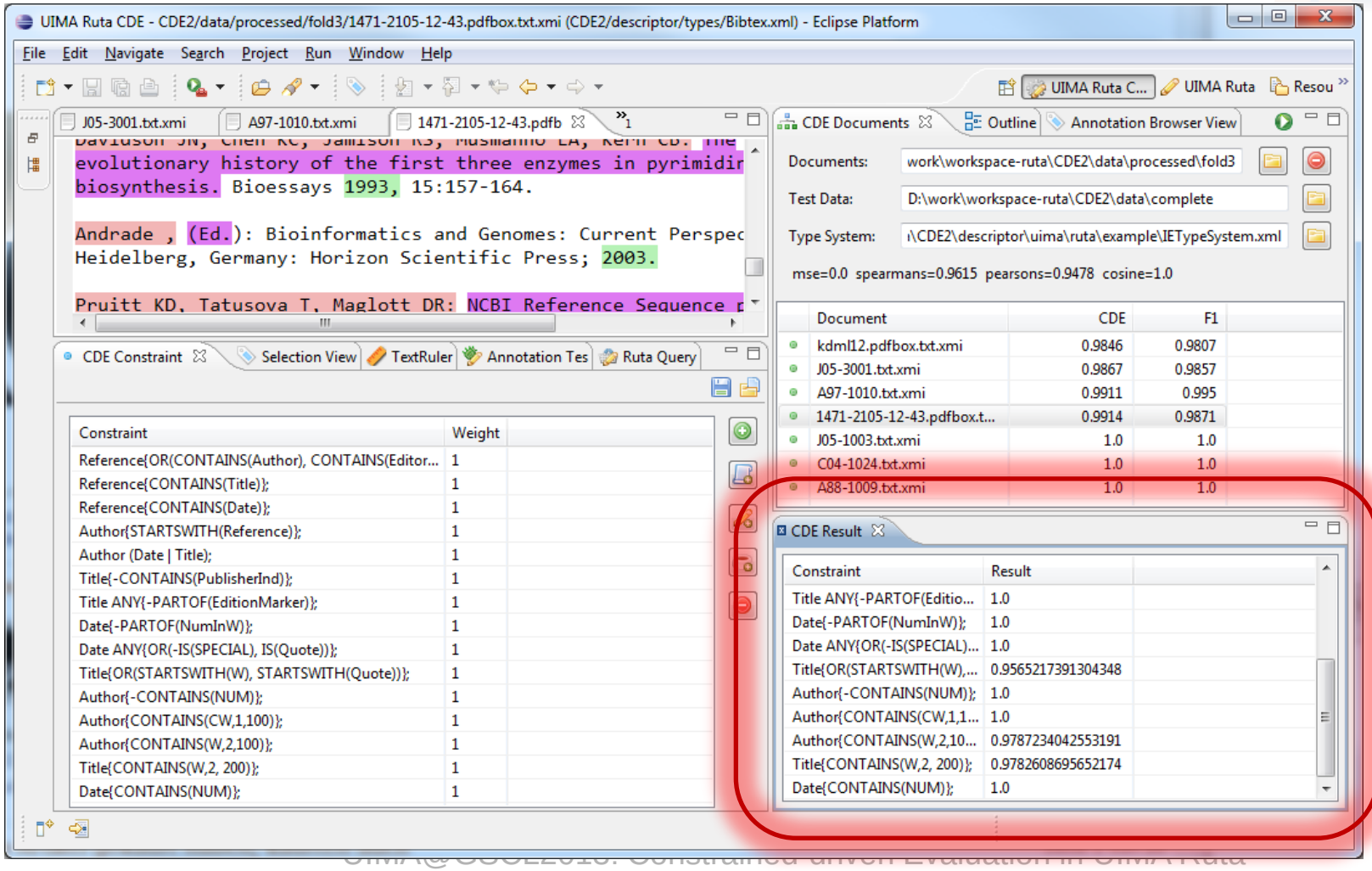
UIMA Ruta CDE Plug-in
- Constraint Specification -



UIMA Ruta CDE Plug-in
- Constraint Development & Results -



UIMA Ruta CDE Plug-in
- Document Details -



Summary: Constraint-driven Evaluation in UIMA Ruta

Predict performance of arbitrary models on unseen data by constraints

- Rule-constraints
- Annotation Distribution constraints

UIMA Ruta workbench plug-in

- Supports rapid prototyping
- and component selection

Experimental study on reference segmentation shows usefulness

Thank you for your attention!

Try CDE in UIMA Ruta 2.1.0:

http://uima.apache.org/ruta.html

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.A., 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.: Condence 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 Articial Intelligence (KI 2010). Springer (2010)

9. Kluegl, P., Toepfer, M., Lemmerich, F., Hotho, A., Puppe, F.: Collective Information Extraction with Context-Specic 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 Classication. 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)