

Automatic Metadata Extraction

The High Energy Physics Use Case

Joseph Boyd

École Polytechnique Fédérale de Lausanne

joseph.boyd@epfl.ch

August 26, 2015

Motivation

- ▶ INSPIRE-HEP digital library at CERN contains over 1 Million documents
- ▶ Manual curation of high energy physics (HEP) papers may be automated with machine learning techniques
- ▶ Custom datasets and specialised features required to model HEP paper characteristics

Aims

Take existing state-of-the-art system for metadata extraction to:

- ▶ demonstrate a qualitative difference between HEP and general papers;
- ▶ propose improvements to model features;
- ▶ run experiments to confirm these improvements, and;
- ▶ draw conclusions about what characterises good feature engineering.

Outline

Introduction

Automatic Metadata Extraction

Data and Features

Key Results

Conclusions

Metadata Extraction

- ▶ *Metadata* refers to content useful to the bibliographic identification of the document
- ▶ *Extraction* refers to the classification of metadata within the document text
- ▶ Several automatic approaches exist: stylistic analysis, knowledge-base, machine learning (CRFs, HMMs, SVMs) ...

Metadata Extraction (Illustration)

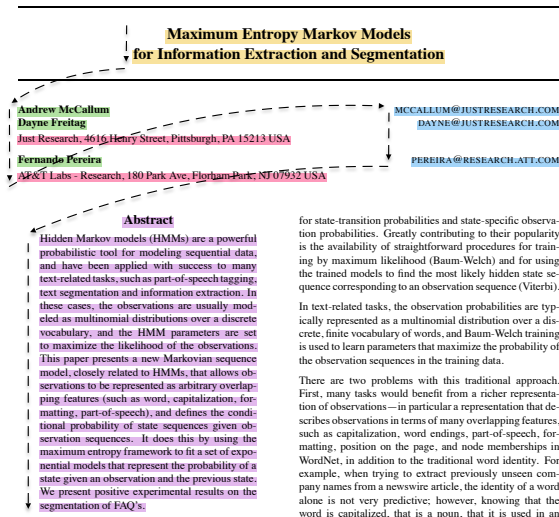


Figure: Tagging of a document header section.

Why Conditional Random Fields?

- ▶ Transition interdependencies implies graphical structure best modelled as a structured sequence
- ▶ Modelling conditional distribution, $p(\mathbf{y}|\mathbf{x})$, sufficient for classification
- ▶ Exploit rich information about observations, \mathbf{x} , without explicitly modelling the underlying probability distribution
- ▶ Classifying metadata may greatly benefit from modelling rich text features (punctuation, font size, layout ...)

GROBID

- ▶ Selected according to performance in study comparing AME systems [2]
- ▶ Open source Java-based tool developed at INRIA, France
- ▶ Manages *cascade* of CRF models for annotating papers in progressively finer detail
- ▶ Uses C++ library *Wapiti* for back-end calculations (model training, prediction)

GROBID - CRF Cascade

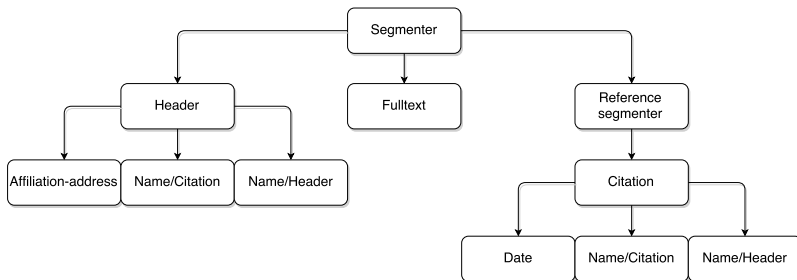


Figure: Cascade of models used by Grobid

Outline

Introduction

Automatic Metadata Extraction

Data and Features

Key Results

Conclusions

Identification of beauty and charm quark jets at LHCb

The LHCb collaboration[†]

Abstract

Identification of jets originating from beauty and charm quarks is important for measuring Standard Model processes and for searching for new physics. The performance of algorithms developed to select b - and c -quark jets is measured using data recorded by LHCb from proton-proton collisions at $\sqrt{s} = 7$ TeV in 2011 and at $\sqrt{s} = 8$ TeV in 2012. The efficiency for identifying a $b(c)$ jet is about 65%(25%) with a probability for misidentifying a light-parton jet of 0.3% for jets with transverse momentum $p_T > 20$ GeV and pseudorapidity $2.2 < \eta < 4.2$. The dependence of the performance on the p_T and η of the jet is also measured.

Submitted to JINST

© CERN on behalf of the LHCb collaboration, license CC-BY-4.0.

Figure: Collaboration field in header section.

HEP Paper Characteristics (ii)

LHCb collaboration

R. Aaij³⁸, B. Adeva³⁷, M. Adinolfi⁴⁶, A. Affolder⁵², Z. Ajaltouni⁵, S. Akar⁶, J. Albrecht⁹, F. Alessio³⁸, M. Alexander⁵¹, S. Ali⁴¹, G. Alkhazov³⁰, P. Alvarez Cartelle⁵³, A.A. Alves Jr⁵⁷, S. Amato², S. Amerio²², Y. Amhis⁷, L. An³, L. Anderlini^{17,g}, J. Anderson⁴⁰, M. Andreotti^{16,f}, J.E. Andrews⁵⁸, R.B. Appleby⁵⁴, O. Aquines Gutierrez¹⁰, F. Archilli³⁸, P. d'Argent¹¹, A. Artamonov³⁵, M. Artuso⁵⁹, E. Aslanides⁶, G. Auremma^{25,n}, M. Baalouch⁵, S. Bachmann¹¹, J.J. Back⁴⁸, A. Badalov³⁶, C. Baesso⁶⁰, W. Baldini^{16,38}, R.J. Barlow⁵⁴, C. Barschel³⁸, S. Barsuk⁷, W. Barter³⁸, V. Batozskaya²⁸, V. Battista³⁹, A. Bay³⁹, L. Beaucourt⁴, J. Beddow⁵¹, F. Bedeschi²³, I. Bediaga¹, L.J. Bel⁴¹, I. Belyaev³¹, E. Ben-Haim⁸, G. Bencivenni¹⁸, S. Benson³⁸, J. Benton⁴⁶, A. Berezhnoy³², R. Bernet⁴⁰, A. Bertolin²², M.-O. Bettler³⁸, M. van Beuzekom⁴¹, A. Bien¹¹, S. Bifani⁴⁵, T. Bird⁵⁴, A. Birnkraut⁹, A. Bizzeti^{17,i}, T. Blake⁴⁸, F. Blanc³⁹, J. Blouw¹⁰, S. Blusk⁵⁹, V. Bocci²⁵, A. Bondar³⁴, N. Bondar^{30,38}, W. Bonivento¹⁵, S. Borghi⁵⁴, M. Borsato⁷, T.J.V. Bowcock⁵², E. Bowen⁴⁰, C. Bozzi¹⁶, S. Braun¹¹, D. Brett⁵⁴, M. Britsch¹⁰, T. Britton⁵⁹, J. Brodzicka⁵⁴, N.H. Brook⁴⁶, A. Bursche⁴⁰, J. Buytaert³⁸, S. Cadeddu¹⁵, R. Calabrese^{16,f}, M. Calvi^{20,k}, M. Calvo Gomez^{36,p}, P. Campana¹⁸, D. Campora Perez³⁸, L. Capriotti⁵⁴, A. Carbone^{14,d}, G. Carboni^{24,l}, R. Cardinale^{19,j}, A. Cardini¹⁵, P. Carniti²⁰, L. Carson⁵⁰, K. Carvalho Akiba^{2,38}, R. Casanova Mohr³⁶, G. Casse⁵², L. Cassina^{20,k}, L. Castillo Garcia³⁸, M. Cattaneo³⁸, Ch. Cauet⁹, G. Cavallero¹⁹, R. Cenci^{23,l}, M. Charles⁸, Ph. Charpentier³⁸, M. Chefdeville⁴, S. Chen⁵⁴, S.-F. Cheung⁵⁵, N. Chiapolini⁴⁰, M. Chrzaszcz⁴⁰, X. Cid Vidal³⁸, G. Ciezarek⁴¹, P.E.L. Clarke⁵⁰, M. Clemencic³⁸, H.V. Cliff⁴⁷, J. Closier³⁸, V. Coco³⁸, J. Cogan⁶, E. Cogneras⁵, V. Cogoni^{15,e}, L. Cojocariu²⁹, G. Collazuoli²², P. Collins³⁸, A. Comerma-Montells¹¹, A. Contu^{15,38}, A. Cook⁴⁶, M. Coombes⁴⁶, S. Coquereau⁸, G. Corti³⁸, M. Corvo^{16,f}, I. Counts⁵⁶, B. Couturier³⁸, G.A. Cowan⁵⁰, D.C. Craik⁴⁸, A. Crocombe⁴⁸, M. Cruz Torres⁶⁰, S. Cunliffe⁵³, R. Currie⁵³, C. D'Ambrosio³⁸, J. Dalseno⁴⁶, P.N.Y. David⁴¹, A. Davis⁵⁷, K. De Bruyn⁴¹, S. De Capua⁵⁴, M. De Cian¹¹, J.M. De Miranda¹, L. De Paula²,

Figure: Collaboration author list.

HEP Paper Characteristics (iii)

- ¹⁸ *Laboratori Nazionali dell'INFN di Frascati, Frascati, Italy*
- ¹⁹ *Sezione INFN di Genova, Genova, Italy*
- ²⁰ *Sezione INFN di Milano Bicocca, Milano, Italy*
- ²¹ *Sezione INFN di Milano, Milano, Italy*
- ²² *Sezione INFN di Padova, Padova, Italy*
- ²³ *Sezione INFN di Pisa, Pisa, Italy*
- ²⁴ *Sezione INFN di Roma Tor Vergata, Roma, Italy*
- ²⁵ *Sezione INFN di Roma La Sapienza, Roma, Italy*
- ²⁶ *Henryk Niewodniczanski Institute of Nuclear Physics Polish Academy of Sciences, Kraków, Poland*
- ²⁷ *AGH - University of Science and Technology, Faculty of Physics and Applied Computer Science, Kraków, Poland*
- ²⁸ *National Center for Nuclear Research (NCBJ), Warsaw, Poland*
- ²⁹ *Horia Hulubei National Institute of Physics and Nuclear Engineering, Bucharest-Magurele, Romania*
- ³⁰ *Petersburg Nuclear Physics Institute (PNPI), Gatchina, Russia*
- ³¹ *Institute of Theoretical and Experimental Physics (ITEP), Moscow, Russia*
- ³² *Institute of Nuclear Physics, Moscow State University (SINP MSU), Moscow, Russia*
- ³³ *Institute for Nuclear Research of the Russian Academy of Sciences (INR RAN), Moscow, Russia*
- ³⁴ *Budker Institute of Nuclear Physics (SB RAS) and Novosibirsk State University, Novosibirsk, Russia*
- ³⁵ *Institute for High Energy Physics (IHEP), Protvino, Russia*
- ³⁶ *Universitat de Barcelona, Barcelona, Spain*
- ³⁷ *Universidad de Santiago de Compostela, Santiago de Compostela, Spain*
- ³⁸ *European Organization for Nuclear Research (CERN), Geneva, Switzerland*
- ³⁹ *Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
- ⁴⁰ *Physik-Institut, Universität Zürich, Zürich, Switzerland*
- ⁴¹ *Nikhef National Institute for Subatomic Physics, Amsterdam, The Netherlands*
- ⁴² *Nikhef National Institute for Subatomic Physics and VU University Amsterdam, Amsterdam, The Netherlands*
- ⁴³ *NSC Kharkiv Institute of Physics and Technology (NSC KIPT), Kharkiv, Ukraine*
- ⁴⁴ *Institute for Nuclear Research of the National Academy of Sciences (KINR), Kyiv, Ukraine*

Figure: Collaboration affiliation list.

HEP Paper Characteristics (iv)

encode different attribute dimensions of an input data space. A good glyph design can enable users to conduct visual search more efficiently during interactive visualization, and facilitate effective learning, memorizing and using the visual encoding scheme. A less effective visual design may suffer from various shortcomings such as being perceptually confusing, semantically ambiguous, difficult to learn and remember, or unable to accommodate low-resolution display devices.

-
- *Eamonn Maguire is with Oxford e-Research Centre and Department of Computer Science, University of Oxford, UK. E-mail: eamonn.maguire@st-annes.ox.ac.uk.*
 - *Philippe Rocca-Serra, Susanna-Assunta Sansone and Min Chen are with Oxford e-Research Centre, University of Oxford, UK. E-mail: {philippe.rocca-serra,susanna-assunta.sansone,min.chen}@oerc.ox.ac.uk.*
 - *Jim Davies is with Department of Computer Science, University of Oxford, UK. E-mail: jim.davies@cs.ox.ac.uk.*

Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.

For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Figure: Discontinuous header data.

Training Data

Two models addressed:

- ▶ *header* model, which classifies *word* tokens as <title>, <author>, <abstract>, etc.
- ▶ *segmentation* model, which classifies *line* tokens as <header>, <body>, <references>, etc.
- ▶ Custom HEP training sets collected for each
- ▶ Customs sets combined with existing CORA datasets during experimentation

Model	HEP	CORA
Header	157 papers	2506 papers
Segmentation	169 papers	125 papers

Table: Number of training instances for each model from each dataset.

Feature Engineering

- Experiments run for different features designed to enhance the models *header* and *segmentation*

Method	Model
Baseline	both
Block Size	header
Character Classes	segmentation
Dictionaries	header
Levenshtein Distance	segmentation
Regularisation	header
Token Extensions	segmentation

Table: Feature engineering experiments

Dictionary Features (*header*)

Dictionaries were derived from the INSPIRE-HEP corpus:

- ▶ affiliations
- ▶ authors
- ▶ collaborations
- ▶ journals
- ▶ titles
- ▶ stop words*

Dictionary-based features were then modelled as,

$$f_{\text{dict}_i}(x_t) = \mathbb{1}_{\{x_t \in \text{dict}_i\}},$$

for each dictionary, dict_i .

Character Class Features (*segmentation*)

Feature functions defined to be,

$$f_{\text{class}_i}(x_t) = \frac{1}{|x_t|} \sum_{n=1}^{|x_t|} \mathbb{1}_{\{x_{ti} \in \text{class}_i\}},$$

for each character class, class_i , where x_t is a token (line), and x_{ti} is the i th character in the line.

Class	Regex
Spacing	$r'[\backslash s]'$
Lower case	$r'[a-z]'$
Upper case	$r'[A-Z]'$
Numeric	$r'[\backslash d]'$
Punctuation	$r'[.,?;:]'$
Special character	$r'^{[\backslash sa-zA-Z d.,?;:]}'$

Table: Character classes used as features.

Levenshtein Distance Features (*segmentation*)

Define similarity function,

$$\text{similarity}(a, b) = 1 - \frac{\text{lev}_{a,b}(|a|, |b|)}{\max(|a|, |b|)}.$$

Then feature function,

$$f_{lev}(x_t) = \begin{cases} 0 & \text{if } 0 \leq \text{similarity}(x_t, x_{t-1}) \leq T_1 \\ 1 & \text{if } T_1 \leq \text{similarity}(x_t, x_{t-1}) \leq T_2 \\ \vdots & \vdots \\ N-1 & \text{if } T_{N-1} \leq \text{similarity}(x_t, x_{t-1}) \leq 1 \end{cases}$$

where T_1, T_2, \dots, T_{N-1} are thresholds selected to create the N categories.

Outline

Introduction

Automatic Metadata Extraction

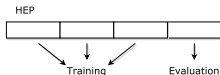
Data and Features

Key Results

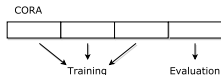
Conclusions

Experiment Setup

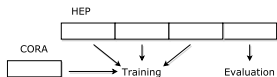
- ▶ 66 experiments run testing combinations of features, model and CV configuration.
- ▶ Models judged primarily on micro average F_1 score, but also with reference to key classes where necessary.



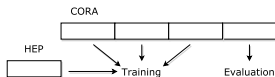
(a) CV HEP



(b) CV CORA



(c) CV HEP append CORA



(d) CV CORA append HEP

Figure: Cross-validation configurations used in experiments.

Header Model (Subsampling CORA dataset)

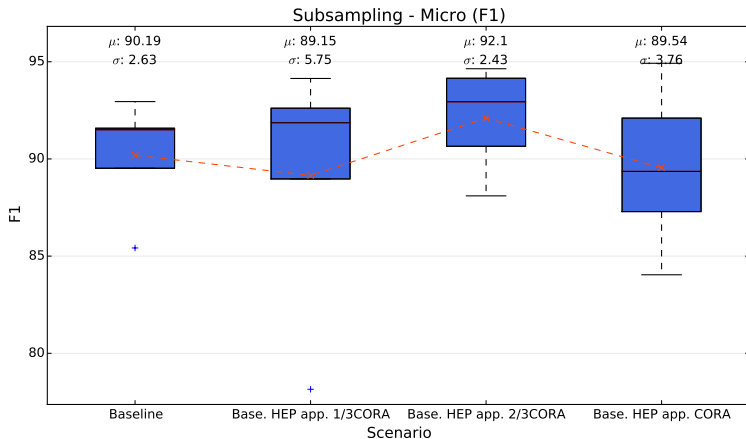


Figure: Appending subsamples of CORA dataset in baseline evaluation.

Header Model (Best Features)

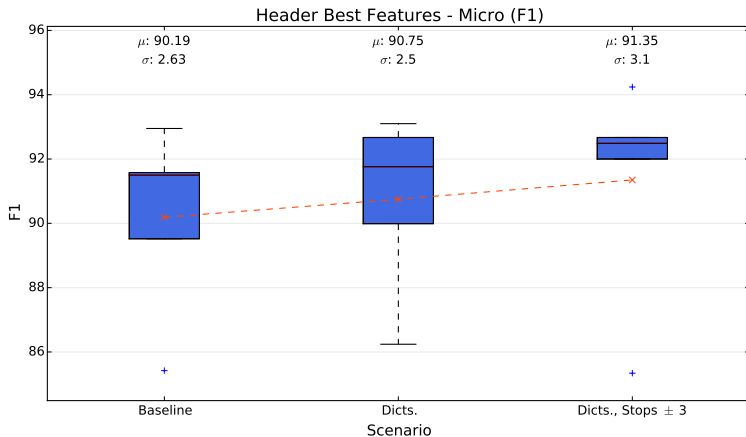


Figure: Best features for header model.

Segmentation Model (Best Features)

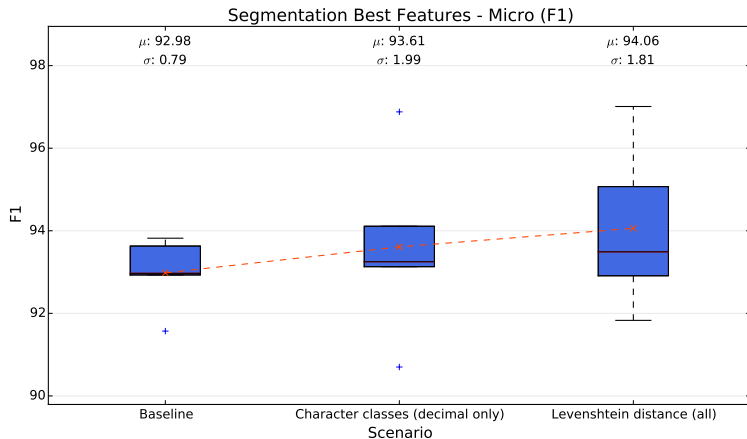


Figure: Best features for segmentation model.

Segmentation Model (Header Field)

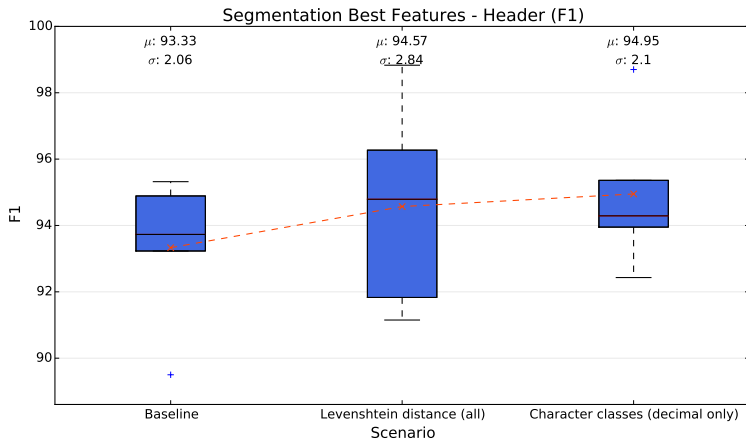


Figure: Best features for segmentation model <header> field.

Segmentation Model (References Field)

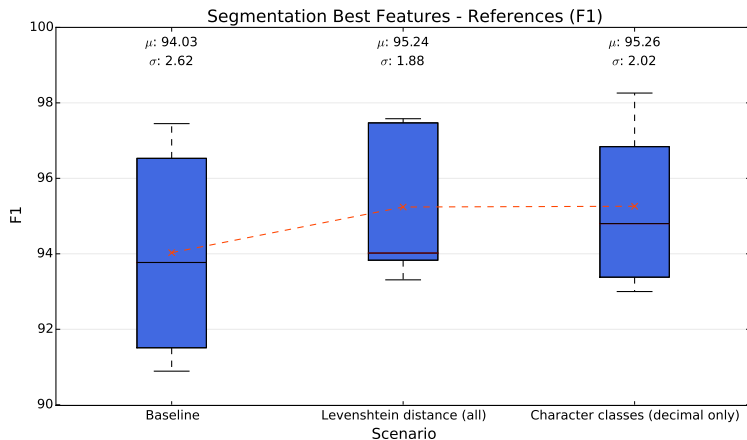


Figure: Best features for segmentation model <references> field.

Segmentation Model (Baseline Confusion Matrix)

Confusion matrix - Segmentation (Baseline, HEP)

acknowledgement	544	5	258			10	1		
annex	13	1805	9078			67	1	3	46
body	142	2601	185466		162	303	45	41	590
cover				259	5	26	1		
footnote	6	8	495		1171	47	42	31	18
header		3	906	43	32	12211	11	6	36
headnote		18	372		56	76	1380	52	35
page	2	10	136		18	12	17	2257	12
references	115		392		11	124	18	5	10871
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references

Figure: Baseline confusion matrix for segmentation model.

Segmentation Model (Character Class Confusion Matrix)

Confusion matrix - Segmentation (Character Classes, HEP)

acknowledgement	616		185			17			
annex	16	1573	9293			3	3	1	124
body	51	2555	185735		87	530	52	28	312
cover				259	5	24	3		
footnote		19	472		1205	39	58	8	17
header			311	1	41	12863	8	3	21
headnote		6	354		55	59	1428	57	30
page		5	116		5	10	22	2297	9
references			433		14	96	21	6	10966
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references

Figure: Confusion matrix for segmentation model with character classes.

Outline

Introduction

Automatic Metadata Extraction

Data and Features

Key Results

Conclusions

Conclusions

- ▶ Qualitative difference between HEP and general papers demonstrated (through inspection, subsampling).
- ▶ Valuable new datasets produced.
- ▶ Successful features offered a dimensionality reduction: dictionaries (12% error reduction), character classes (24% and 21% on <header> and <references> classifications).



R. Aaij, B. Adeva, M. Adinolfi, A. Affolder, Z. Ajaltouni, S. Akar, J. Albrecht, F. Alessio, M. Alexander, S. Ali, et al. Identification of beauty and charm quark jets at LHCb. *arXiv preprint arXiv:1504.07670*, 2015.



M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 385–386. ACM, 2013.



E. Maguire, P. Rocca-Serra, S.-A. Sansone, J. Davies, and M. Chen. Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2603–2612, 2012.