Automatic Metadata Extraction
The High Energy Physics Use Case

Joseph Boyd

Supervisors: Dr. Martin Rajman, Dr. Gilles Louppe

École Polytechnique Fédérale de Lausanne

joseph.boyd@epfl.ch

September 1, 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 new features to enhance models;
- run experiments to test these new features, and;
- draw conclusions about what characterises good feature engineering.

Outline

Metadata Extraction

- *Metadata* refers to content useful to the bibliographpic 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)...

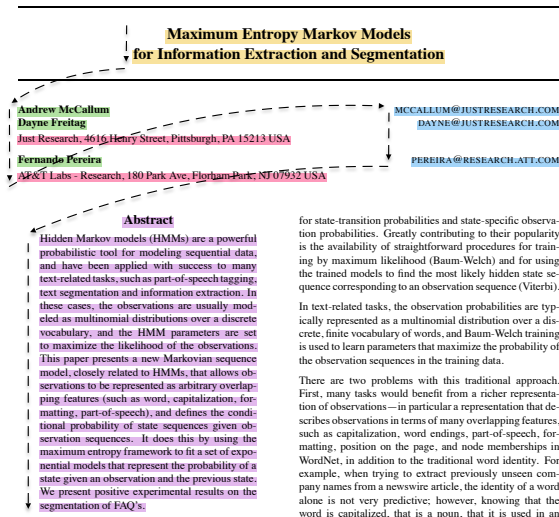# Metadata Extraction (Illustration)

Figure: Tagging of a document header section.

Mathematical Formulation

A CRF factorises in the following as,

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{\sum_{y'} p(\mathbf{x}, \mathbf{y}')} = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_{k} \lambda_k F_k(y_t, y_{t-1}, x_t) \right\},$$

- $Z(\mathbf{x}) = \sum_{y'} \exp\left\{ \sum_{k} \lambda_{ij} F_k(y_t', y_{t-1}', x_t) \right\}$, known as the partition function.
- $F_k(\mathbf{x}, y) = \sum_{t}^{T} f_k(\mathbf{x}, y)$, where f_k is a function expressing a feature.
- It is in choosing the form of the functions, $f(\cdot)$, explicitly that we perform feature engineering.

Why Conditional Random Fields?

- Transition interdependencies implies graphical structure $\rightarrow$ model 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
- Open source Java-based tool developed at INRIA
- Manages *cascade* of CRF models for annotating papers in progressively finer detail
- Uses C++ library *Wapiti* for back-end calculations (model training, prediction)
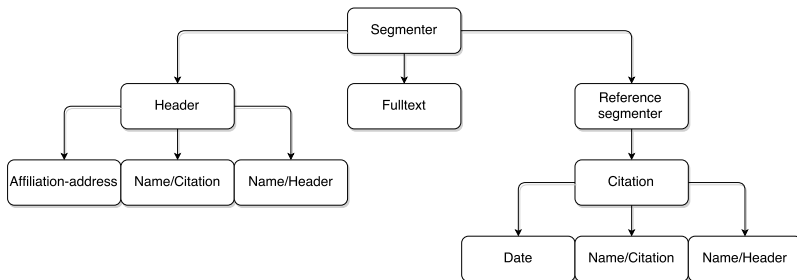
GROBID - CRF Cascade

Figure: Cascade of models used by GROBID

Outline

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 $\sqrt{s} = 7\,\text{TeV}$ in 2011 and at $\sqrt{s} = 8\,\text{TeV}$ 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 $p_\text{T} > 20\,\text{GeV}$ and pseudorapidity $2.2 < \eta < 4.2$. The dependence of the performance on the p_T and η of the jet is also measured.

Figure: Collaboration field in header section.

HEP Paper Characteristics (ii)

LHCb collaboration

R. Aaij[38], B. Adeva[37], M. Adinolfi[46], A. Affolder[52], Z. Ajaltouni[5], S. Akar[6], J. Albrecht[9],
F. Alessio[38], M. Alexander[51], S. Ali[41], G. Alkhazov[30], P. Alvarez Cartelle[53], A.A. Alves Jr[57],
S. Amato[2], S. Amerio[22], Y. Amhis[7], L. An[3], L. Anderlini[17,g], J. Anderson[40], M. Andreotti[16,f],
J.E. Andrews[58], R.B. Appleby[54], O. Aquines Gutierrez[10], F. Archilli[38], P. d'Argent[11],
A. Artamonov[35], M. Artuso[59], E. Aslanides[6], G. Auriemma[25,n], M. Baalouch[5], S. Bachmann[11],
J.J. Back[48], A. Badalov[36], C. Baesso[60], W. Baldini[16,38], R.J. Barlow[54], C. Barschel[38],
S. Barsuk[7], W. Barter[38], V. Batozskaya[28], V. Battista[39], A. Bay[39], L. Beaucourt[4], J. Beddow[51],
F. Bedeschi[23], I. Bediaga[1], L.J. Bel[41], I. Belyaev[31], E. Ben-Haim[8], G. Bencivenni[18], S. Benson[38],
J. Benton[46], A. Berezhnoy[32], R. Bernet[40], A. Bertolin[22], M.-O. Bettler[38], M. van Beuzekom[41],
A. Bien[11], S. Bifani[45], T. Bird[54], A. Birnkraut[9], A. Bizzeti[17,i], T. Blake[48], F. Blanc[39],
J. Blouw[10], S. Blusk[59], V. Bocci[25], A. Bondar[34], N. Bondar[30,38], W. Bonivento[15], S. Borghi[54],
M. Borsato[7], T.J.V. Bowcock[52], E. Bowen[40], C. Bozzi[16], S. Braun[11], D. Brett[54], M. Britsch[10],
T. Britton[59], J. Brodzicka[54], N.H. Brook[46], A. Bursche[40], J. Buytaert[38], S. Cadeddu[15],
R. Calabrese[16,f], M. Calvi[20,k], M. Calvo Gomez[36,p], P. Campana[18], D. Campora Perez[38],
L. Capriotti[54], A. Carbone[14,d], G. Carboni[24,l], R. Cardinale[19,j], A. Cardini[15], P. Carniti[20],
L. Carson[50], K. Carvalho Akiba[2,38], R. Casanova Mohr[36], G. Casse[52], L. Cassina[20,k],
L. Castillo Garcia[38], M. Cattaneo[38], Ch. Cauet[9], G. Cavallero[19], R. Cenci[23,t], M. Charles[8],
Ph. Charpentier[38], M. Chefdeville[4], S. Chen[54], S.-F. Cheung[55], N. Chiapolini[40], M. Chrzaszcz[40],
X. Cid Vidal[38], G. Ciezarek[41], P.E.L. Clarke[50], M. Clemencic[38], H.V. Cliff[47], J. Closier[38],
V. Coco[38], J. Cogan[6], E. Cogneras[5], V. Cogoni[15,e], L. Cojocariu[29], G. Collazuol[22], P. Collins[38],
A. Comerma-Montells[11], A. Contu[15,38], A. Cook[46], M. Coombes[46], S. Coquereau[8], G. Corti[38],
M. Corvo[16,f], I. Counts[56], B. Couturier[38], G.A. Cowan[50], D.C. Craik[48], A. Crocombe[48],
M. Cruz Torres[60], S. Cunliffe[53], R. Currie[53], C. D'Ambrosio[38], J. Dalseno[46], P.N.Y. David[41],
A. Davis[57], K. De Bruyn[41], S. De Capua[54], M. De Cian[11], J.M. De Miranda[1], L. De Paula[2],

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.*

Figure: Discontinuous header data.

Training Data

Two models addressed:

- *header* model, which classifies *word* tokens as <title>, <author>, , etc.
- *segmentation* model, which classifies *line* tokens as <header>, <body>, <references>, etc.
- Custom HEP training sets collected for each
- HEP dataset 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*
- 66 experiments run testing combinations of features, model and CV configuration.

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 *ith* character in the line.

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

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

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

Outline

Header Model (Best Features)
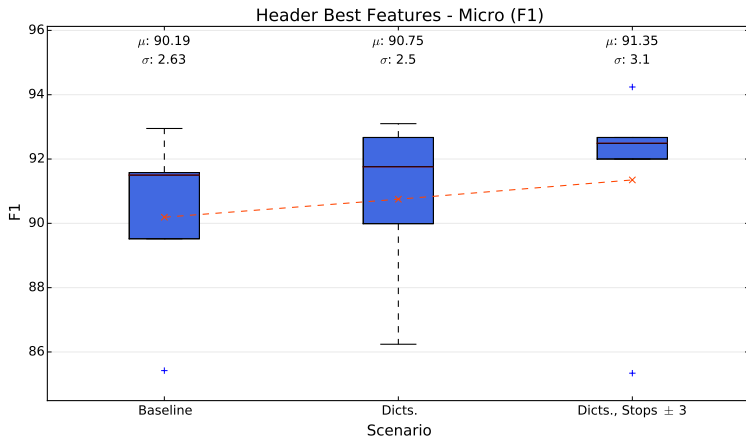
Figure: Best features for header model.

Segmentation Model (Header Field)
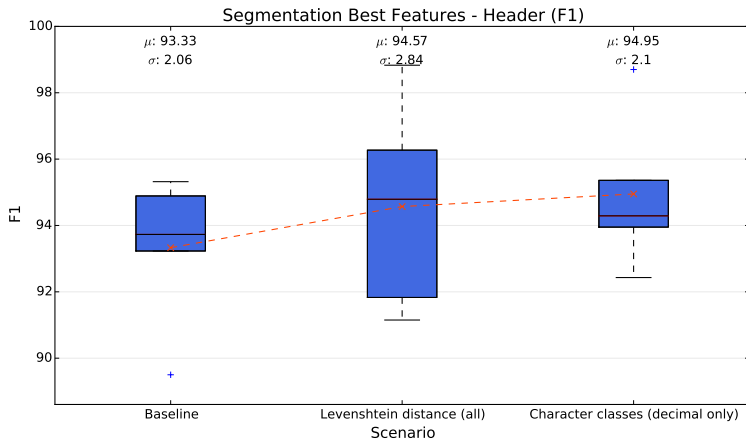
Figure: Best features for segmentation model <header> field.
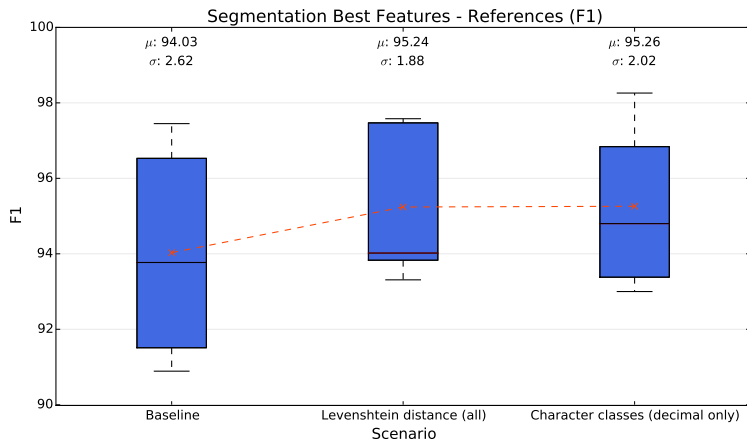
Segmentation Model (References Field)

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

Segmentation Model (Baseline Confusion Matrix)

Figure: Baseline confusion matrix for segmentation model.

Segmentation Model (Character Class Confusion Matrix)

Figure: Confusion matrix for segmentation model with character classes.

Outline

Conclusions

- Qualitative difference between HEP and general papers demonstrated (through inspection, subsampling).
- New datasets produced.
- Performance-enhancing features found for both HEP and general cases.
- Successful features yielded error reductions: dictionaries (12% error reduction), character classes (24% and 21% on <header> and <references> classifications).