Automatic Metadata Extraction
The High Energy Physics Use Case

Joseph Boyd

École Polytechnique Fédérale de Lausanne

joseph.boyd@epfl.ch

August 30, 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:

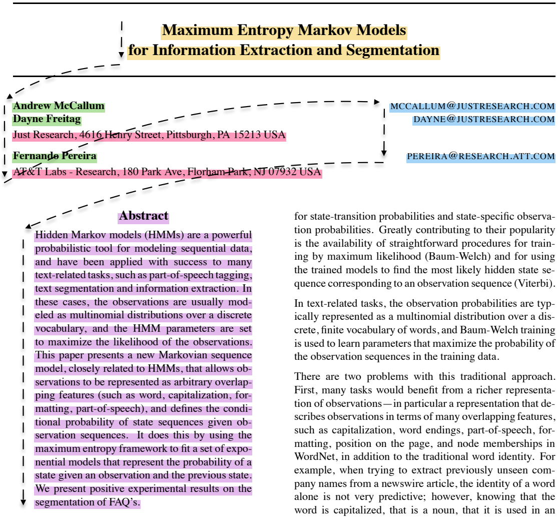
- demonstrate a qualitative difference between HEP and general papers;
- propose new features to enhance models;
- run experiments to test these new features, and;
- draw conclusions about what characterises good feature engineering.

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)



Maximum Entropy Markov Models for Information Extraction and Segmentation

Andrew McCallum MCCALLUM@JUSTRESEARCH.COM
Dayne Freitag DAYNE@JUSTRESEARCH.COM
Just Research, 4616 Henry Street, Pittsburgh, PA 15213 USA

Fernando Pereira PEREIRA@RESEARCH.ATT.COM
AT&T Labs - Research, 180 Park Ave, Florham Park, NJ 07932 USA

Abstract

Hidden Markov models (HMMs) are a powerful probabilistic tool for modeling sequential data, and have been applied with success to many text-related tasks, such as part-of-speech tagging, text segmentation and information extraction. In these cases, the observations are usually modeled as multinomial distributions over a discrete vocabulary, and the HMM parameters are set to maximize the likelihood of the observations. This paper presents a new Markovian sequence model, closely related to HMMs, that allows observations to be represented as arbitrary overlapping features (such as word, capitalization, formatting, part-of-speech), and defines the conditional probability of state sequences given observation sequences. It does this by using the maximum entropy framework to fit a set of exponential models that represent the probability of a state given an observation and the previous state. We present positive experimental results on the segmentation of FAQ's.

for state-transition probabilities and state-specific observation probabilities. Greatly contributing to their popularity is the availability of straightforward procedures for training by maximum likelihood (Baum-Welch) and for using the trained models to find the most likely hidden state sequence corresponding to an observation sequence (Viterbi).

In text-related tasks, the observation probabilities are typically represented as a multinomial distribution over a discrete, finite vocabulary of words, and Baum-Welch training is used to learn parameters that maximize the probability of the observation sequences in the training data.

There are two problems with this traditional approach. First, many tasks would benefit from a richer representation of observations—in particular a representation that describes observations in terms of many overlapping features, such as capitalization, word endings, part-of-speech, formatting, position on the page, and node memberships in WordNet, in addition to the traditional word identity. For example, when trying to extract previously unseen company names from a newswire article, the identity of a word alone is not very predictive; however, knowing that the word is capitalized, that is a noun, that it is used in an

Figure: Tagging of a document header section.

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, 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

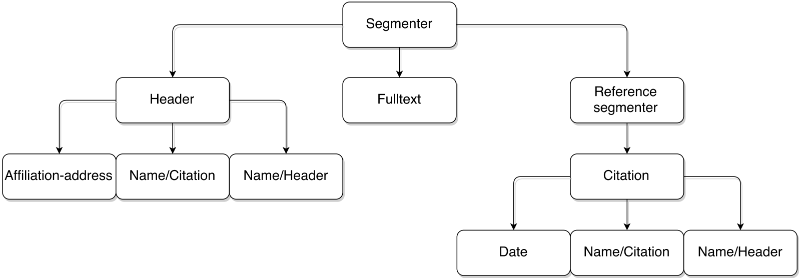


Figure: Cascade of models used by GROBID

Outline

HEP Paper Characteristics (i)

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 $\sqrt{s} = 7\,\mathrm{TeV}$ in 2011 and at $\sqrt{s} = 8\,\mathrm{TeV}$ 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 $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ 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

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.*

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>, <refernces>, 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,

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

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‘[ˆ\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,

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

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

Outline

Header Model (Best Features)

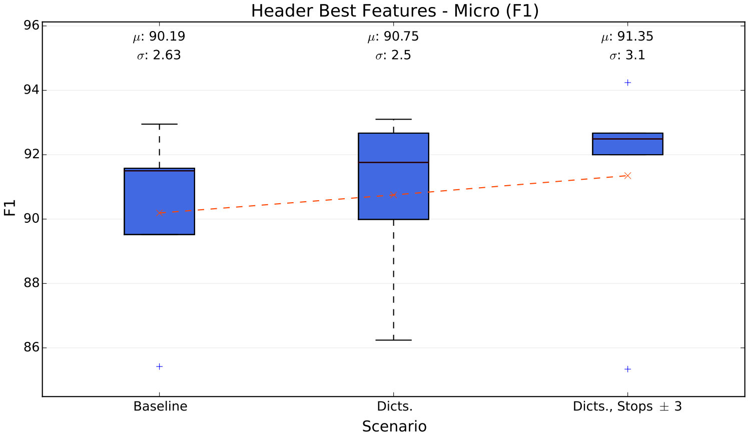


Figure: Best features for header model.

Segmentation Model (Header Field)

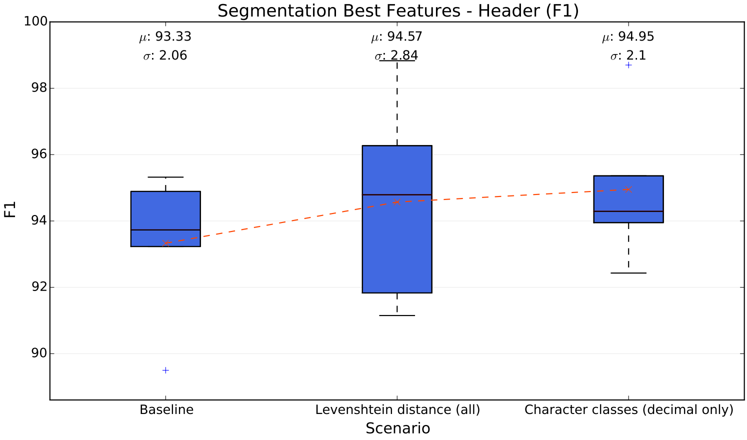


Figure: Best features for segmentation model <header> field.

Segmentation Model (References Field)

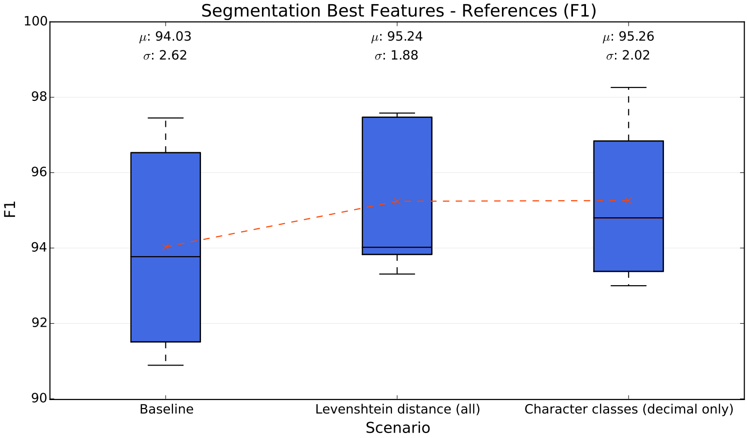


Figure: Best features for segmentation model <references> field.

Segmentation Model (Baseline Confusion Matrix)

Confusion matrix - Segmentation (Baseline, HEP)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
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

Figure: Baseline confusion matrix for segmentation model.

Segmentation Model (Character Class Confusion Matrix)

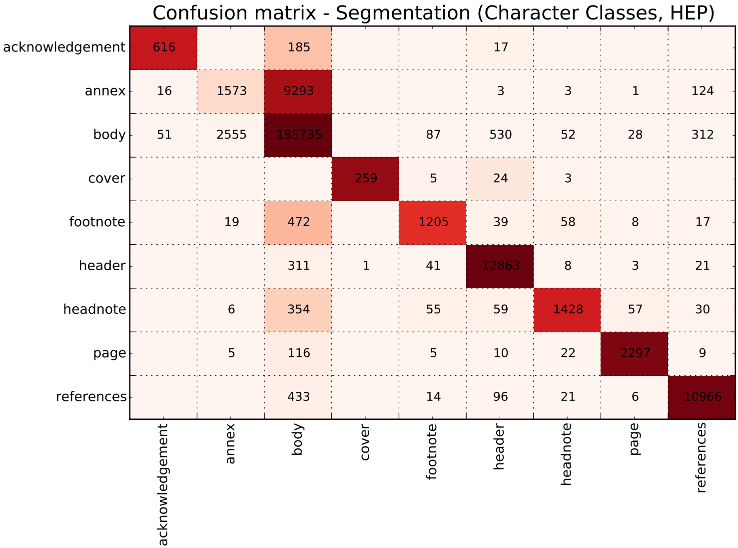


Figure: Confusion matrix for segmentation model with character classes.

Outline

Conclusions

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