Automatic Metadata Extraction with Conditional Random Fields

Joseph Boyd

École Polytechnique Fédérale de Lausanne

joseph.boyd@epfl.ch

April 13, 2015

Outline

The Problem

- Metadata extraction of scientific articles is a well-studied problem
- Metadata usually refers to header information (title, authors, publisher, abstract...), bibliographic information (citations and their sub-components), document structure (sections, etc.)
- Extraction of metadata refers to processing and then *classifying* text blocks and other components of the document.
- The problem is difficult because of the wide variety of document formats that may appear

The Problem - An Illustration

Accurate Information Extraction from Research Papers using Conditional Random Fields

Fuchun Peng
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
fuchun@cs.umass.edu

Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
mccallum@cs.umass.edu

Abstract

With the increasing use of research paper search engines, such as CiteSeer, for both literature search and hiring decisions, the accuracy of such systems is of paramount importance. This paper employs Conditional Random Fields (CRFs) for the task of extracting various common fields from the headers and citation of research papers. The basic theory of CRFs is becoming well-understood, but best-practices for applying them to real-world data requires additional exploration. This paper makes an empirical exploration of several factors, including variations on Gaussian, exponential and hyperbolic-L_1 priors for improved regularization, and several classes of features and Markov order. On a standard benchmark data set, we achieve new state-of-the-art perfor-

Previous work in information extraction from research papers has been based on two major machine learning techniques. The first is hidden Markov models (HMM) (Seymore et al., 1999; Takasu, 2003). An HMM learns a generative model over input sequence and labeled sequence pairs. While enjoying wide historical success, standard HMM models have difficulty modeling multiple non-independent features of the observation sequence. The second technique is based on discriminatively-trained SVM classifiers (Han et al., 2003). These SVM classifiers can handle many non-independent features. However, for this sequence labeling problem, Han et al. (2003) work in a two stages process: first classifying each line independently to assign it label, then adjusting these labels based on an additional classifier that examines larger windows of labels. Solving the information extraction problem in two steps looses the tight interaction between state transitions and observations.

Hindawi Publishing Corporation
Advances in High Energy Physics
Volume 2015, Article ID 292767, 10 pages
http://dx.doi.org/10.1155/2015/292767

Hindawi

Research Article

Quintessence and Holographic Dark Energy in $f(T)$ Gravity

M. Zubair

Department of Mathematics, COMSATS Institute of Information Technology, Lahore 54000, Pakistan

Correspondence should be addressed to M. Zubair; mzubairkk@gmail.com

Received 22 September 2014; Revised 21 December 2014; Accepted 2 January 2015

Academic Editor: Filipe R. Joaquim

We regard $f(T)$ theory as an efficient tool to explain the current cosmic acceleration and associate its evolution with the known dark energy models. The numerical scheme is applied to reconstruct $f(T)$ theory from dark energy model with constant equation of state parameter and holographic dark energy model. We set the model parameters ω_ϑ and c as describing the different evolution eras and show the distinctive behavior of each case realized in $f(T)$ theory. We also present the future evolution of reconstructed $f(T)$ and find that it is consistent with the recent observations.

The Problem - Some Assumptions

- We are always dealing with PDF documents, and an OCR tool gives us access to the plaintext and OCR information (font, size, orientation,...).
- Metadata has structure and is not completely random, but it is infeasible to model deterministically $\Rightarrow$ the task is inherently error-prone.
- There is no one-size-fits-all model for processing a full document $\Rightarrow$ the problem must be decomposed.

The Problem - Solutions

There are a number of approaches to metadata extraction:

- Template-based
- Knowledge base (consult online repositories)
- Machine learning techniques (HMMs, SVMs, CRFs)

In practice, tools for metadata extraction combine these approaches (an example of which we will see later).

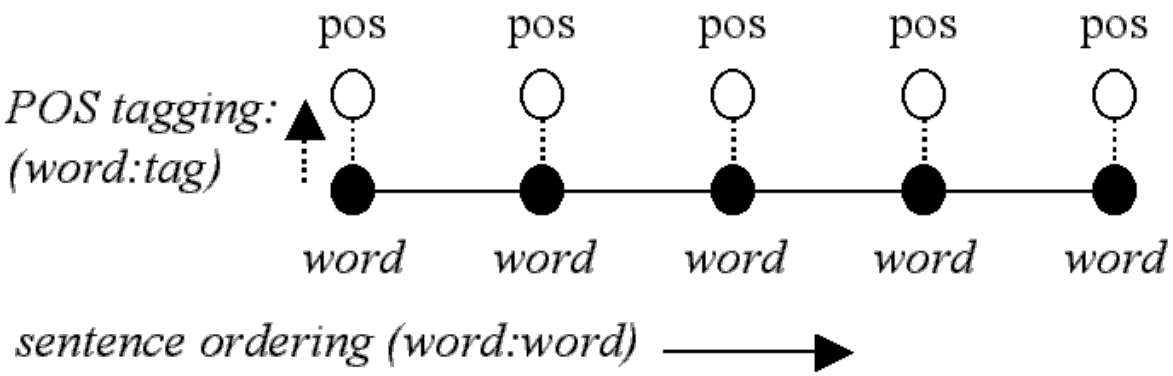
Outline

Hidden Markov Models (HMMs)

- Hidden Markov Models (HMM) form a probabilistic model on sequential data.
- We have a sequence of observations, and we want to predict the "hidden" states which generate them.
- The Markov condition postulates a dependence of one state to the next as we move in time.
- HMMs have applications in fields as diverse as text processing, bioinformatics, and artificial intelligence.

Hidden Markov Models (HMMs) - Example

A well-known application of HMMs is from Natural Language Processing (NLP) - Part of Speech (PoS) tagging.



Metadata extraction is a similar problem...

Hidden Markov Models (HMMs) - Prediction

- A Hidden Markov Model factorises in the following way:

$$p(\mathbf{x}, \mathbf{y}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$

- The probabilities $p(y_t|y_{t-1})$ are "transition" probabilities, and $p(x_t|y_t)$"emission" probabilities, and must be calculated in advance.
- When it comes to prediction, we use dynamic programming Viterbi algorithm ($\mathcal{O}(T|S|^2)$) to maximise the conditional distribution:

$$\mathbf{y}_{prediction} = \underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t) \right\}$$

Take home message: we can (efficiently) predict the hidden sequence by maximising a likelihood.

Outline

Logistic Regression

- A logistic regression is used for classifying a data sample into two (binary) or more (multi) categories, thus,

$$\hat{y}_{prediction} = \sigma\left(\boldsymbol{\beta}^T \cdot \boldsymbol{x}_{sample}\right),$$

where $\hat{y}$ is the prediction (represented as a probability), $\boldsymbol{x} = [x_0, x_1, ..., x_D]^T$ is a data sample, and $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_D]^T$ is the vector of parameters we must *learn*

- We construct a (maximum log likelihood) cost function in terms of this parameter vector,

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{n=1}^{N} y_n \boldsymbol{\beta}^T \boldsymbol{x}_n - \log[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_n)]$$

Solving a Logistic Regression

- Building a regression model is equivalent to solving a convex optimisation problem (i.e. minimising the cost function)
- We know the form of the model, and we have a set of (training) data
- We want to choose the model parameters for which the error is minimised (think line of best fit)
- We use a numerical method to find the global minimum of error, for example, the method of gradient descent:

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^{k} - \alpha \nabla \mathcal{L}(\boldsymbol{\beta}^{k})$$

Take home message: we can automatically build mathematical functions for making predictions.

Outline

Conditional Random Fields (CRFs)

- CRFs belong with HMMs to a class of models called *graphical* models.
- Classifiers such as logistic regression only predict a single class; graphical models predict a full sequence.
- CRFs learn paramaters like logistic regression, unlike HMMs where they are precomputed.
- Modelling the conditional probability, $p(\mathbf{y}|\mathbf{x})$ is sufficient for classification, so a CRF avoids modelling the distribution on $\mathbf{x}$, while allowing a freer choice of features (expressed as feature functions).
- Whereas with a HMM we model an observation's *occurrence*, with a CRF we can model richer information about an observation.

Conditional Random Fields (CRFs)

- A conditional random field factorises as,

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{\sum_{y'} p(\mathbf{x}, \mathbf{y}')},$$

where

$$p(\mathbf{x}, \mathbf{y}) = \exp\left\{ \sum_{i \in S} \sum_{j \in S} \lambda_{ij} F_{ij}(\mathbf{y}) + \sum_{i \in S} \sum_{o \in O} \mu_{io} F_{io}(\mathbf{x}, \mathbf{y}) \right\}.$$

Choosing $F_{io}(\mathbf{x}, \mathbf{y}) \approx$ feature engineering

Conditional Random Fields (CRFs)

- Like a logistic regression we have a cost function to minimise using greatest ascent techniques and solve for model parameters.
- The optimisation algorithm is usually an approximation (l-BFGS).
- There are some further tricks during training (inference algorithms).
- Prediction is (as with HMMs) done using the Viterbi algorithm, since our factorisation is a function of both observations ($\mathbf{x}$) and states ($\mathbf{y}$).

Take home message: CRFs are very similar to HMMs: we predict in the same way, but we learn parameters through training. Modelling the conditional distribution gives us the freedom to exploit richer features.

Outline

Grobid

- Grobid (GeneRation Of BIbliographic Data) is a Java-based tool for managing CRF models for metadata extraction.
- It coordinates the training and usage of a "cascade" of models, computed with a CRF engine backend.
- It may also query the web service "CrossRef" for verification.

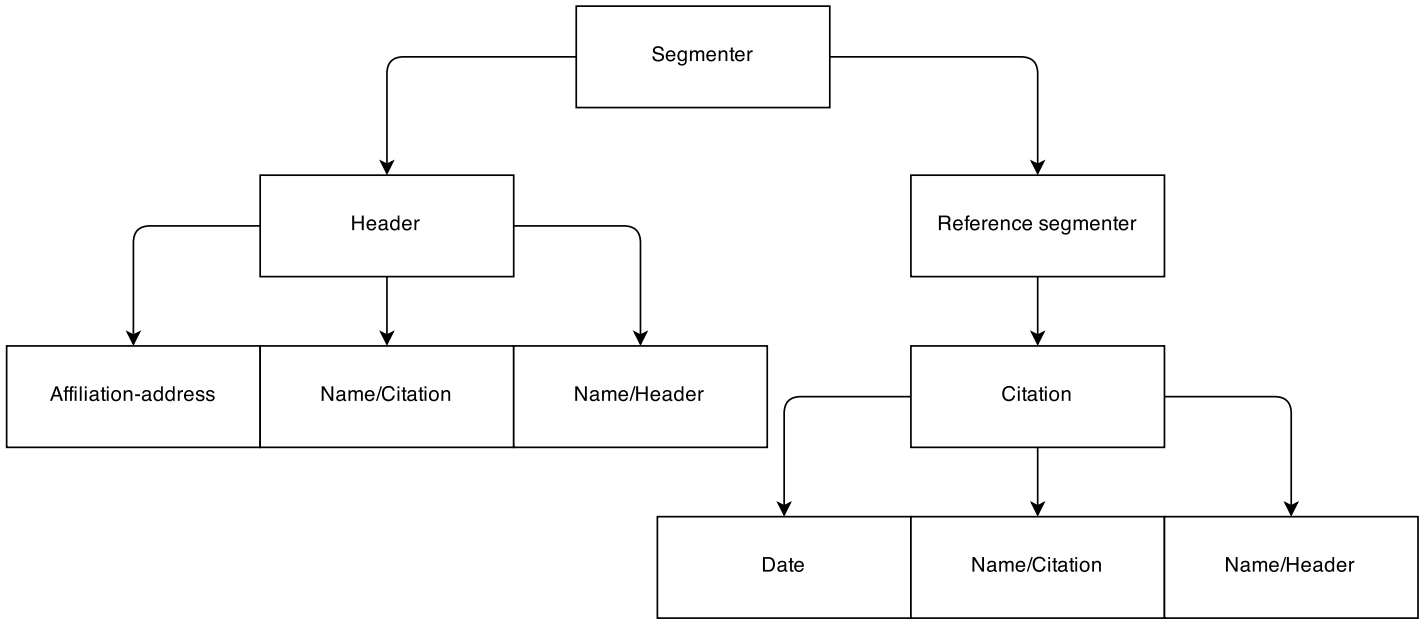


Figure : Cascade of models used by Grobid

Grobid - Details

```
<bibl>
<author>V. Gundelach and D. Eisenburger</author>, \&
    quot;
<title level="a">Principle of a direction sensitive
    borehole antenna with advanced technology and data
     examples</title>, \&quot; in
<title level="m">Proceedings of the 4th International
    Workshop on Advanced Ground Penetrating Radar (
    IWAGPR \&apos;07)</title>, pp.
<biblScope type="pp">28âĂŞ31</biblScope>,
<date>June 2007</date>.
</bibl>
```

Figure : Sample tagged citation for Grobid training input.

Grobid - Training and Prediction

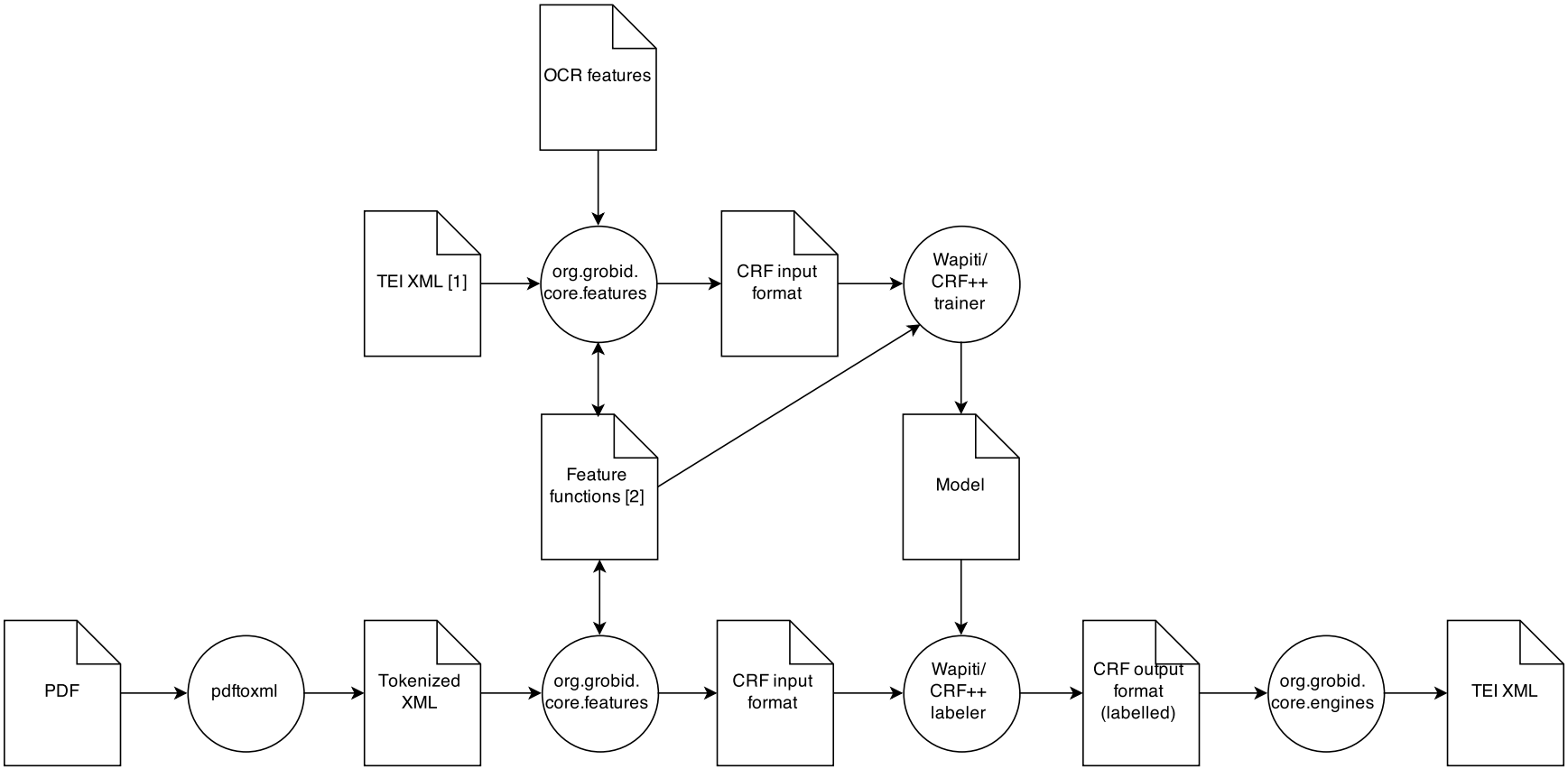


Figure : Training and prediction with Grobid

Outline

Initial Results - Reference Segmentation

label	accuracy	precision	recall	f1
<label>	99.78	99.56	100	99.78
<reference>	100	100	100	100
(micro average)	99.89	99.78	100	99.89
(macro average)	99.89	99.78	100	99.89

Table : Evaluation results for reference segmentation

Initial Results - Citations

label	accuracy	precision	recall	f1
<author>	92.97	49.07	91.31	63.84
<journal>	97.06	78.61	68.69	73.32
<volume>	97.72	97.57	63.25	76.75
<title>	95.93	61.96	91.51	73.89
<pages>	96.5	67.71	87.66	76.41
<date>	95.99	74.38	72.93	73.65
<booktitle>	98.79	60	9.02	15.69
<location>	98.89	0	0	0
<issue>	99.47	0	0	0
<note>	99.77	0	0	0
<editor>	99.77	0	0	0
<publisher>	99.37	0	0	0
<web>	99.97	0	0	0
<tech>	99.85	0	0	0
<institution>	99.82	0	0	0
<pubnum>	99.96	0	0	0
(micro average)	98.24	66.23	71.71	68.86
(macro average)	98.24	30.58	30.27	28.35

Table : Evaluation results for citations

Outline

Project Objectives

- HEP papers have some particularities, such as collaborations of thousands of authors.
- We aim to create tailored feature functions that improve performance (probably focusing on the header and segmentation).
- As well as to configure Grobid for live use.

References


LIPINSKI, M., YAO, K., BREITINGER, C., BEEL, J., AND GIPP, B. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In JCDL (2013), pp. 385-386.


C. Sutton and A. McCallum (2006), An Introduction to Conditional Random Fields for Relational Learning., In, In Introduction to Statistical Relational Learning., MIT Press.