Automatic Metadata Extraction with Conditional Random Fields

Joseph Boyd

École Polytechnique Féqérale de Lausanne

joseph.boyd@epfl.ch

April 12, 2015

Outline

The Problem

- Metadata extraction of scientific articles is a difficult and well-studied problem
- Metadata usually refers to header information (title, authors, affiliations, publisher...), 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 - 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

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_9 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 this 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)

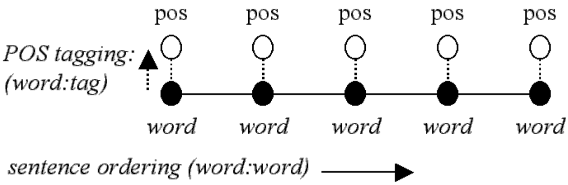
Outline

Hidden Markov Models (HMMs)

- Hidden Markov Models (HMM) build 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,

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

  where $\hat{y}$ is the prediction (represented as a probability), $\mathbf{x} = [x_0, x_1, ..., x_D]^T$ 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,

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

Solving a Logistic Regression

- Building a regression model is equivalent to solving a convex optimisation problem (i.e. maximising 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 broader class of models called *graphical* models
- Classifiers such as logistic regression only predict a single class; graphical models predict a full sequence
- Because modelling the conditional probability, $p(\mathbf{y}|\mathbf{x})$ is sufficient for classification, a CRF avoids modelling the distribution on $\mathbf{x}$, while allowing freedom over the choice of features (expressed as feature functions)
- Whereas with a HMM we model the observation of word tokens, with a CRF, we can also model richer information about a token

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

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

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

Conditional Random Fields (CRFs)

- Like a logistic regression we have a cost function to minimise using greatest ascent techniques and solve for model parameters
- There are some further tricks during training (inference algorithms)
- And the optimisation algorithm is usually an approximation (l-BFGS)
- But in the end, because the cost is dependent on the labels, we use the Viterbi algorithm to do predictions, just like HMMs.

Take home message: CRFs are very similar to HMMs; we predict in the same way; but we learn the parameters through training like logistic regression; and modelling the conditional distribution gives us the freedom of exploiting richer features.

Outline

1. The Problem
2. Theory
   - Hidden Markov Models
   - Logistic Regression
   - Conditional Random Fields
3. Grobid
4. Initial Results
5. Next Steps

Grobid

- Grobid (GeneRation Of BIbliographic Data) is a Java-based tool for managing CRF models
- It coordinates the training and usage of a “cascade” of models, computed with a CRF engine backend
- CROSSREF

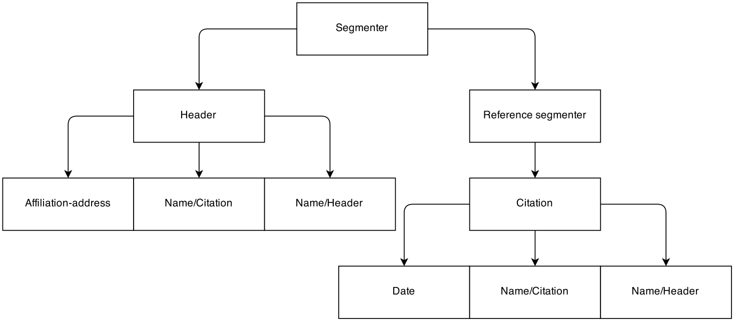


Figure : Cascade of models used by Grobid

Grobid - Details

TEI example

Outline

Initial Results

TABLE HERE

Outline

1. The Problem
2. Theory
   - Hidden Markov Models
   - Logistic Regression
   - Conditional Random Fields
3. Grobid
4. Initial Results
5. Next Steps

Next Steps

Project Objectives

- There are some special cases we want to deal with in HEP papers
- Take the existing state-of-the-art in metadata extraction and optimise it for HEP papers
- We will use Conditional Random Fields (CRFs) to build a probabilistic model
- CRFs combine ideas from log linear models (logistic regressions) and hidden Markov models (HMMs)

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.