Automatic Metadata Extraction: The High Energy Physics Use Case

Joseph Boyd

July 1, 2015

Contents

1 Introduction

1.1 Motivation

1.2 Aims

1.3 Main Results

1.4 Outline

2 Supervised Sequence Learning

In this section we present the state-of-the-art technique for metadata extraction, conditional random fields (CRF). For completeness, we include a background history of related machine learning techniques and their optimisation algorithms. Notably, we pinpoint the piece of the mathematical model relevant to our work on feature engineering. Finally, we describe Wapiti, a general-purpose "engine" for training and applying CRF models.

2.1 Log-linear Models

2.2 Graphical Models

2.2.1 Hidden Markov Models

2.3 Conditional Random Fields

2.3.1 Feature Engineering

2.3.2 Wapiti

Lack of support for numeric features imposes constraints on the choice of features possible. Any numeric-based idea must be discretised

3 Automatic Metadata Extraction

3.1 Metadata Extraction

3.2 Related Work

3.3 GROBID

[Show here Grobid vs. refextract]

4 Implementation and Data

4.1 Extensions

4.2 Data Acquisition

5 Results and Analysis

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if min(i, j)} = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & \text{otherwise} \end{cases}$$

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

5.1 Experiment Setup

Months of CPU time? (parallelised), 64 experiments (before an combination experiments are run) Mind you, though we aren't explicitly interested in identifying headnotes, footnotes, page numbers etc., correctly classifying them does spare the important categories (header, references) from garbage data.

5.2 Evaluation Method

5.3 Baseline

5.3.1 Header model - Cora dataset

5.3.2 Header model - Cora dataset appending HEP dataset

5.3.3 Header model - Cora and HEP combined datasets

5.3.4 Header model - HEP dataset

5.3.5 Header model - HEP dataset appending CORA dataset

5.3.6 Header model - HEP dataset appending 1/3 CORA dataset

5.3.7 Header model - HEP dataset appending 2/3 CORA dataset

5.3.8 Segmentation model - Cora dataset

5.3.9 Segmentation model - Cora dataset appending HEP dataset

5.3.10 Segmentation model - Cora and HEP combined datasets

5.3.11 Segmentation model - HEP dataset

5.3.12 Segmentation model - HEP dataset appending CORA dataset

5.4 Regularisation

5.4.1 Header model - $L2 = 0$

5.4.2 Header model - $L2 = 1e^{-6}$

5.4.3 Header model - $L2 = 1e^{-5}$

5.4.4 Header model - $L2 = 1e^{-4}$

5.4.5 Header model - $L2 = 1e^{-3}$

5.5 Dictionaries

5.5.1 Header model - HEP dataset

5.5.2 Header model - HEP dataset appending CORA dataset

5.5.3 Segmentation model - HEP dataset

5.5.4 Segmentation model - HEP dataset appending CORA dataset

5.5.5 Header Model - HEP dataset - 2^{nd} Degree Features

5.5.6 Header Model - HEP dataset Appending CORA - 2^{nd} Degree Features

5.5.7 Header Model - HEP dataset - 3^{rd} Degree Features

5.5.8 Header Model - HEP dataset Appending CORA - 3^{rd} Degree Features

5.6 Dictionaries + stop words

7 References

8 Appendices