Automatic Metadata Extraction: The High Energy Physics Use Case

Joseph Boyd

June 25, 2015

Contents

1 Introduction **3**
1.1 Motivation . . . 3
1.2 Aims . . . 3
1.3 Main Results . . . 3
1.4 Outline . . . 3

2 Supervised Sequence Learning **3**
2.1 Log-linear Models . . . 3
2.2 Graphical Models . . . 3
2.2.1 Hidden Markov Models . . . 3
2.3 Conditional Random Fields . . . 3
2.3.1 Feature Engineering . . . 3
2.3.2 Wapiti . . . 3

3 Automatic Metadata Extraction **3**
3.1 Metadata Extraction . . . 3
3.2 Related Work . . . 3
3.3 GROBID . . . 3

4 Implementation and Data **3**
4.1 Extensions . . . 3
4.2 Data Acquisition . . . 3

5 Results and Analysis **3**
5.1 Experiment Setup . . . 3
5.2 Evaluation Method . . . 3
5.3 Baseline . . . 3
5.3.1 Header model - Cora dataset . . . 3
5.3.2 Header model - Cora dataset appending HEP dataset . . . 3
5.3.3 Header model - Cora and HEP combined datasets . . . 3
5.3.4 Header model - HEP dataset . . . 3
5.3.5 Header model - HEP dataset appending CORA dataset . . . 3
5.3.6 Header model - HEP dataset appending 1/3 CORA dataset . . . 3
5.3.7 Header model - HEP dataset appending 2/3 CORA dataset . . . 3
5.3.8 Segmentation model - Cora dataset . . . 3
5.3.9 Segmentation model - Cora dataset appending HEP dataset . . . 3
5.3.10 Segmentation model - Cora and HEP combined datasets . . . 3
5.3.11 Segmentation model - HEP dataset . . . 3
5.3.12 Segmentation model - HEP dataset appending CORA dataset . . . 3

5.4 Regularisation . . . 3
5.4.1 Header model - $L2 = 0$ . . . 3
5.4.2 Header model - $L2 = 1e^{-6}$ . . . 3
5.4.3 Header model - $L2 = 1e^{-5}$ . . . 3
5.4.4 Header model - $L2 = 1e^{-4}$ . . . 3
5.4.5 Header model - $L2 = 1e^{-3}$ . . . 3
5.5 Dictionaries . . . 3
5.5.1 Header model - HEP dataset . . . 3
5.5.2 Header model - HEP dataset appending CORA dataset . . . 3
5.5.3 Segmentation model - HEP dataset . . . 3
5.5.4 Segmentation model - HEP dataset appending CORA dataset . . . 3
5.5.5 Header Model - HEP dataset - 2^{nd} Degree Features . . . 3
5.5.6 Header Model - HEP dataset Appending CORA - 2^{nd} Degree Features . . . 3
5.5.7 Header Model - HEP dataset - 3^{rd} Degree Features . . . 3
5.5.8 Header Model - HEP dataset Appending CORA - 3^{rd} Degree Features . . . 3
5.6 Dictionaries + stop words . . . 3
5.6.1 Header model - HEP dataset . . . 3
5.6.2 Header model - HEP dataset appending CORA dataset . . . 3
5.6.3 Segmentation model - HEP dataset . . . 3
5.6.4 Segmentation model - HEP dataset appending CORA dataset . . . 3
5.6.5 Header Model - HEP dataset - 2^{nd} Degree Features . . . 3
5.6.6 Header Model - HEP dataset Appending CORA - 2^{nd} Degree Features . . . 3
5.6.7 Header Model - HEP dataset - 3^{rd} Degree Features . . . 3
5.6.8 Header Model - HEP dataset Appending CORA - 3^{rd} Degree Features . . . 3
5.7 Token Selection . . . 3
5.7.1 Segmentation Model - HEP dataset - 5 Tokens . . . 3
5.7.2 Segmentation Model - HEP dataset - 10 Tokens . . . 3
5.7.3 Segmentation Model - HEP dataset - 15 Tokens . . . 3
5.7.4 Segmentation Model - HEP dataset - 20 Tokens . . . 3
5.8 Levenshtein . . . 3
5.8.1 Segmentation Model - HEP dataset - Binary Threshold (0.05) . . . 3
5.8.2 Segmentation Model - HEP dataset - Binary Threshold (0.1) . . . 3
5.8.3 Segmentation Model - HEP dataset - Binary Threshold (0.2) . . . 3
5.8.4 Segmentation Model - HEP dataset - Binary Threshold (0.4) . . . 3
5.8.5 Segmentation Model - HEP dataset - Binary Threshold (0.8) . . . 3
5.8.6 Segmentation Model - HEP dataset - Ternary Threshold . . . 3
5.8.7 Segmentation Model - HEP dataset - Quaternary Threshold . . . 3
5.9 Line Shape . . . 3
5.9.1 Segmentation Model - HEP dataset - Binary Threshold . . . 3
5.9.2 Segmentation Model - HEP dataset - Ternary Threshold . . . 3
5.10 Template Matching . . . 3
5.10.1 Segmentation Model - HEP dataset . . . 3

6 Conclusion **3**
6.1 Summary . . . 3
6.1.1 Key Results . . . 3
6.2 Future Work . . . 3

7 References **3**

8 Appendices **3**

1 Introduction

1.1 Motivation

1.2 Aims

1.3 Main Results

1.4 Outline

2 Supervised Sequence Learning

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

3 Automatic Metadata Extraction

3.1 Metadata Extraction

3.2 Related Work

3.3 GROBID

4 Implementation and Data

4.1 Extensions

4.2 Data Acquisition

5 Results and Analysis

5.1 Experiment Setup

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