

## MASTER'S THESIS

### Information extraction for on-line job advertisements

Au, Kwok Chung

*Date of Award:*  
2004

[Link to publication](#)

#### **General rights**

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

# Information Extraction for On-line Job Advertisements

AU Kwok Chung

A thesis submitted in partial fulfillment of the requirements

for the degree of

Master of Philosophy

Principal Supervisor: Dr. CHEUNG Kwok Wai

Hong Kong Baptist University

June 2004

# Abstract

The Web has widely been used as a low-cost and yet effective way to disseminate various kind of information, where job advertisements (ads) is one of the typical examples. Its rapid growth has quickly resulted in the need for some efficient ways to locate and analyze relevant information from heterogeneous data sources. The commonly used keyword search engines often return more information than needed. In order to support more precise information search using semantics, information extraction systems have been proposed for extracting some specific target items from on-line documents as their semantic tags.

On-line job ads normally contain grammatical, telegraphic as well as ungrammatical text. Also, their writing styles and layout structures are relatively informal. These characteristics make the pattern matching approach become one of the suitable candidates for extracting job ads related information. In the first part of this thesis, we propose a rule-based system for information extraction by matching patterns featured by lexical and typesetting information. The extraction rules were derived automatically via induction. The experimental results show that the proposed system can choose the correct rule representation via the proposed induction process and successfully extract company names from on-line job ads up to an accuracy of 92.13%.

Within a job ad, fields other than “company name” normally do not carry typesetting attributes and extracting them has to be solely based on words. Modeling a large set of word sequences via generalization from a small set of training data has been known to be

non-trivial. In the second part of this thesis, we adopt the *hidden Markov model* (HMM) — a probabilistic model for modeling temporal processes, as a flexible pattern matcher for information extraction. While most of the related works require the HMM topology to be constructed manually, we take the state-merging approach instead and propose an algorithm for learning the HMM topology. The proposed learning process is governed by a localized version of the Kullback Leibler divergence. Extensive experiments have been performed to evaluate the performance of the proposed learning algorithm. A set of job advertisements obtained from a newsgroup was used for the evaluation. The resulting HMM can achieve a recall rate up to 83.33% at a precision of 62.50%.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 On-line Document Analysis . . . . .	2
1.1.1 Text Segmentation . . . . .	2
1.1.2 Information Extraction . . . . .	3
1.2 The Automatic Training Approach . . . . .	5
1.3 Thesis Organization . . . . .	6
<b>2 Background</b>	<b>7</b>

2.1	Literature Review on Information Extraction . . . . .	8
2.1.1	Information Extraction From Grammatical Free Text . . . . .	8
2.1.2	Information Extraction From Telegraphic Text . . . . .	9
2.1.3	Information Extraction Using Wrappers . . . . .	10
2.1.4	Automatic Rule Generation . . . . .	10
2.2	Hidden Markov Models . . . . .	12
2.2.1	Introduction . . . . .	12
2.2.2	Applying HMM to IE . . . . .	14
<b>3</b>	<b>A Rule-Based Information Extraction System</b>	<b>16</b>
3.1	Introduction . . . . .	17
3.2	Representation . . . . .	18
3.2.1	Token Classes . . . . .	19
3.2.2	Word Lists . . . . .	20
3.2.3	Rule Modeling . . . . .	21
3.3	Rule Induction . . . . .	21
3.3.1	Rule Learning . . . . .	22
3.3.2	Pattern Extraction . . . . .	28
3.3.3	False Alarms Rejection . . . . .	28
3.3.4	Applying Prior Knowledge via Exclusion Lists . . . . .	29
3.4	Experiments . . . . .	29
3.4.1	Preprocessing . . . . .	30

3.4.2	Extraction of Company Names . . . . .	31
3.5	Conclusion . . . . .	33
<b>4</b>	<b>A Hidden Markov Model for With Adaptive Structure Information Extraction</b>	<b>38</b>
4.1	Introduction . . . . .	39
4.2	Background . . . . .	40
4.2.1	Hidden Markov Model . . . . .	40
4.2.2	HMM for Information Extraction . . . . .	42
4.2.3	Learning HMM Topology . . . . .	43
4.3	KL Divergence based State Merging . . . . .	46
4.3.1	Formulation . . . . .	46
4.3.2	KL Divergence Between HMMs . . . . .	47
4.3.3	The Proposed Learning Algorithm . . . . .	49
4.3.4	Smoothing of HMM Parameters . . . . .	52
<b>5</b>	<b>Implementation Details And Experimental Results</b>	<b>55</b>
5.1	Experimental Results . . . . .	56
5.1.1	Setup . . . . .	56
5.1.2	Data Pre-processing . . . . .	56
5.1.3	Learning Process Speed-up . . . . .	57
5.1.4	Performance Evaluation . . . . .	57
5.1.5	Determining $\alpha$ . . . . .	58

5.1.6	Setting Smoothing Parameters . . . . .	59
5.1.7	Incorporating Prior Knowledge . . . . .	60
5.2	Two Baseline Algorithms for Overall Performance Comparison . . . . .	63
5.2.1	Keyword Matching . . . . .	63
5.2.2	Baseline HMM . . . . .	64
5.3	Overall Performance Comparison . . . . .	64
5.3.1	Comparison Between Keyword Matching Methods . . . . .	64
5.3.2	Extracting Job Titles . . . . .	65
5.3.3	Performance For Extracting Job Specification: “application”, “language”, “platform” . . . . .	66
<b>6</b>	<b>Conclusion and Future Work</b>	<b>70</b>
6.1	Summary of Thesis . . . . .	71
6.1.1	A Rule-based Information Extraction System . . . . .	71
6.1.2	An HMM with Adaptive Structure for Information Extraction . . . . .	71
6.1.3	Rule-Based System Versus Adaptive HMM System . . . . .	71
6.2	Future Research Directions . . . . .	72
6.2.1	Further Performance Comparison for the Rule-based System . . . . .	72
6.2.2	Hybrid Approach for Topology Learning . . . . .	72
6.2.3	Performance boosting strategies . . . . .	73
6.2.4	Further Speedup in State Merging . . . . .	74
6.2.5	Computing KL Divergence Between HMMs . . . . .	74



<b>A</b>	<b>Datasets Used in Experiments</b>	<b>75</b>
A.1	RAPIER’s dataset . . . . .	75
<b>B</b>	<b>Computation of KL Divergence Between Stochastic Grammars</b>	<b>78</b>
B.1	Preliminaries . . . . .	78
B.2	Entropy of a SDRL . . . . .	79
B.3	ALGERIA . . . . .	83
B.4	MDI . . . . .	84
<b>C</b>	<b>The Complexity of the localized KL-Divergence HMM Topology Learning Algorithm</b>	<b>85</b>
	<b>Curriculum Vitae</b>	<b>93</b>