Unsupervised Learning in Natural Language Processing

Proceedings of the Workshop

Edited by
Andrew Kehler and Andreas Stolcke

21 June 1999
University of Maryland
College Park, Maryland, USA

Published by the Association for Computational Linguistics

Unsupervised Learning in Natural Language Processing

Proceedings of the Workshop

Edited by
Andrew Kehler and Andreas Stolcke

21 June 1999
University of Maryland
College Park, Maryland, USA

Published by the Association for Computational Linguistics

Order additional copies from:

Association for Computational Linguistics
75 Paterson Street, Suite 9
New Brunswick, NJ 08901 USA
+1-732-342-9100 phone
+1-732-342-9339 fax
acl@aclweb.org

PREFACE

Many of the successes achieved from using learning techniques in natural language processing (NLP) have utilized the supervised paradigm, in which models are trained from data annotated with the target concepts to be learned. While it is worthwhile to utilize annotated data when it is available, one might argue that the future success of learning for natural language systems cannot depend on a paradigm requiring that large, annotated data sets be created for each new problem or application. The costs of annotation are prohibitively time and expertise intensive, and the resulting corpora may be too susceptible to restriction to a particular domain, application, or genre. Thus, long-term progress in NLP is likely to be dependent on the use of unsupervised and weakly supervised learning techniques, which do not require large annotated data sets. The Workshop on Unsupervised Learning in Natural Language Processing, held on June 21st, 1999 at the University of Maryland, College Park, Maryland, USA, was organized around the goals of discussing, promoting, and presenting new research results (positive and negative) in the use of such methods in NLP. The workshop was sponsored by the Association for Computational Linguistics (ACL), and endorsed by SIGNLL, the ACL Special Interest Group on Natural Language Learning.

This volume contains the nine papers out of twenty-one submissions that were accepted for presentation at the workshop. The program also includes two invited talks, by Michael Brent (Johns Hopkins University) and Lillian Lee (Cornell University). The workshop will conclude with a panel session on unsupervised learning in NLP with the following participants: Eric Brill (Johns Hopkins University), John Lafferty (Carnegie-Mellon University), Andrew McCallum (Carnegie-Mellon University and Just Research), and Janyce Wiebe (New Mexico State University).

We would like to thank all authors who showed their interest by submitting papers to the workshop. We would also like to thank those members of the program committee and others who contributed to the reviewing process: Steve Abney (AT&T Laboratories), Eric Brill (Johns Hopkins University), Rebecca Bruce (University of North Carolina at Asheville), Eugene Charniak (Brown University), Michael Collins (AT&T Laboratories), Marie desJardins (SRI International), Moises Goldszmidt (SRI International), John Lafferty (Carnegie-Mellon University), Lillian Lee (Cornell University), Chris Manning (University of Sydney), Ray Mooney (University of Texas, Austin), Srini Narayanan (ICSI, Berkeley), Fernando Pereira (AT&T Laboratories), David Powers (Flinders University of South Australia), Adwait Ratnaparkhi (IBM Research), Dan Roth (University of Illinois at Urbana-Champaign), Richard Sproat (AT&T Laboratories), Janyce Wiebe (New Mexico State University), and David Yarowsky (Johns Hopkins University).

Andrew Kehler and Andreas Stolcke
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

WORKSHOP PROGRAM

9:00-9:15 WELCOME

9:15-10:15 INVITED TALK

Unsupervised Segmentation of Japanese
Lillian Lee

10:15-10:35 COFFEE BREAK

10:40-12:10 PAPER SESSION I

Dual Distributional Verb Sense Disambiguation with Small Corpora and Machine Readable Dictionaries
Jeong-Mi Cho, Jungyun Seo, and Gil Chang Kim

A Computational Approach to Deciphering Unknown Scripts
Kevin Knight and Kenji Yamada

Resolving Translation Ambiguity using Non-parallel Bilingual Corpora
Genichiro Kikui

12:10-1:30 LUNCH

1:30-3:00 PAPER SESSION II

Detecting Sub-Topic Correspondence through Bipartite Term Clustering
Zvika Marx, Ido Dagan, and Eli Shamir

Text Classification by Bootstrapping with Keywords, EM and Shrinkage
Andrew McCallum and Kamal Nigam

Hiding a Semantic Hierarchy in a Markov Model
Steven Abney and Marc Light

3:00-4:00 INVITED TALK

Combinatoric Generative Models: Bridging the Gap from Minimum Description Length to Complete Probability Models
Michael Brent

4:00-4:20 COFFEE BREAK

4:20-5:50 PAPER SESSION III

The Applications of Unsupervised Learning to Japanese Grapheme-Phoneme Alignment
Timothy Baldwin and Hozumi Tanaka

Unsupervised Lexical Learning with Categorial Grammars
Stephen Watkinson and Suresh Manandhar

Unsupervised Learning of Derivational Morphology from Inflectional Lexicons
Éric Gaussier

5:50-6:30 PANEL SESSION

Eric Brill, Johns Hopkins University
John Lafferty, Carnegie-Mellon University
Andrew McCallum, Carnegie-Mellon University and Just Research
Janyce Wiebe, New Mexico State University

TABLE OF CONTENTS