ACL-05

Feature Engineering for Machine Learning in Natural Language Processing

Proceedings of the Workshop

29 June 2005
University of Michigan
Ann Arbor, Michigan, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Sponsorship gratefully received from
Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Introduction

The ACL 2005 Workshop on Feature Engineering for Machine Learning in Natural Language Processing is an opportunity to explore the various dimensions of feature engineering for problems that are of interest to the ACL community. Feature Engineering encompasses feature design, feature selection, feature induction, studies of feature impact (including feature ablation studies), and related topics. In 2003, there was a NIPS workshop on feature engineering ("Feature Extraction and Feature Selection"), but the focus was not on NLP problems specifically. Also, although the various aspects of feature engineering have been dealt with at times in various ACL forums, until now, to our knowledge, the spotlight has never been shone directly on this topic specifically for NLP and language technology problems. We feel that now is the time to look more closely.

As experience with machine learning for solving natural language processing tasks accumulates in the field, practitioners are finding that feature engineering is as critical as the choice of machine learning algorithm, if not more so. Feature engineering significantly affects the performance of systems and deserves greater attention. Also, in the wake of the shift in our field away from knowledge engineering and of the successes of data-driven and statistical methods, researchers are likely to make further progress by incorporating additional, sometimes familiar, sources of knowledge as features. Feature design may benefit from expert insight even where the relative merits of features must be assessed through empirical techniques from data. Although some experience in the area of feature engineering is to be found in the theoretical machine learning community, the particular demands of natural language processing leave much to be discovered.

In the call for papers, we expressed our intent of bringing together practitioners of NLP, machine learning, information extraction, speech processing, and related fields with the goal of sharing experimental evidence for successful approaches to feature engineering. Judging by the quality and diversity of the submissions received, we believe we have succeeded, and the resulting program should be of great interest to many researchers in the ACL community. We hope that the workshop will contribute to the distillation of best practices and to the discovery of new sources of knowledge and features previously untapped.

We also extend an open invitation to the reader to continue investigation in all aspects of feature engineering for machine learning in NLP, including:

- Novel methods for discovering or inducing features, such as mining the web for closed classes, useful for indicator features.
- Comparative studies of different feature selection algorithms for NLP tasks.
- Error analysis tools that help researchers to identify ambiguous cases that could be disambiguated by the addition of features.
- Error analysis of various aspects of feature induction, selection, representation.
- Issues with representation, e.g., strategies for handling hierarchical representations, including decomposing to atomic features or by employing statistical relational learning.

- Techniques used in fields outside NLP that prove useful in NLP.
- The impact of feature selection and feature design on such practical considerations as training time, experimental design, domain independence, and evaluation.
- Analysis of feature engineering and its interaction with specific machine learning methods commonly used in NLP.
- Ensemble methods employing diverse types of features.
- Studies of methods for inducing a feature set, for example by iteratively expanding a base feature set.
- Issues with representing and combining real-valued and categorical features for NLP tasks.

We anticipate that contributions in these areas will move the field of NLP and language technologies forward, with greater system performance and further insight into our own data and perhaps language itself.

We wish to thank all of the researchers who submitted papers to the workshop. Also, thanks go to the entire program committee (see next page) and those who assisted them in their reviewing responsibilities.

Best regards,

Eric Ringger, Microsoft Research (USA)

20 May 2005

Organizer:

Eric Ringger, Microsoft Research (USA)

Program Committee:

Eric Ringger, Microsoft Research (USA)
Simon Corston-Oliver, Microsoft Research (USA)
Kevin Duh, University of Washington (USA)
Matthew Richardson, Microsoft Research (USA)
Oren Etzioni, University of Washington (USA)
Andrew McCallum, University of Massachusetts at Amherst (USA)
Dan Bikel, IBM Research (USA)
Olac Fuentes, INAOE (Mexico)
Christopher Manning, Stanford University (USA)
Kristina Toutanova , Stanford University (USA)
Hideki Isozaki, NTT Communication Science Laboratories (Japan)
Caroline Sporleder, University of Edinburgh (UK)

Additional Reviewers:

Thamar Solorio, INAOE (Mexico)

Invited Speaker:

Andrew McCallum, University of Massachusetts at Amherst (USA)

Table of Contents

Conference Program

Wednesday, June 29, 2005

8:45–9:00 Opening Remarks

Session W4.1: Classification

A Novel Machine Learning Approach for the Identification of Named Entity Relations
Tianfang Yao and Hans Uszkoreit

Feature Engineering and Post-Processing for Temporal Expression Recognition Using Conditional Random Fields
Sisay Fissaha Adafre and Maarten de Rijke

Temporal Feature Modification for Retrospective Categorization
Robert Liebscher and Richard K. Belew

10:30–11:00 Break

11:00–12:00 Invited Talk by Andrew McCallum

Using Semantic and Syntactic Graphs for Call Classification
Dilek Hakkani-Tür, Gokhan Tur and Ananlada Chotimongkol

12:30–14:00 Lunch

Session W4.2: Discourse and Syntax

Feature-Based Segmentation of Narrative Documents
David Kauchak and Francine Chen

Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns
Adriane Boyd, Whitney Gegg-Harrison and Donna Byron

Engineering of Syntactic Features for Shallow Semantic Parsing
Alessandro Moschitti, Bonaventura Coppola, Daniele Pighin and Roberto Basili

15:30–16:00 Break

Wednesday, June 29, 2005 (continued)

Session W4.3: Feature Sources

Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms
Michael Gamon and Anthony Aue

Studying Feature Generation from Various Data Representations for Answer Extraction
Dan Shen, Geert-Jan M. Kruijff and Dietrich Klakow