Proceedings of the

2004 Conference on Empirical Methods in Natural Language Processing

A meeting of
SIGDAT, a Special Interest Group of the ACL
held in conjunction with ACL 2004

25-26 July 2004
Barcelona, Spain

Preface

EMNLP 2004 represents the twelfth year of SIGDAT meetings, and reflects an unprecedented new scale that will doubtless impact heavily the future course to be charted by the community. We received a record 257 submissions, compared with approximately 140 in the two preceding years. In order to avoid dropping the 24% acceptance rate of preceding years, a record 58 papers were accepted (of which 2 were subsequently withdrawn) while 17 were committed to other conferences. This volume is over 450 pages—yet another record.

Even with fairly long days, two scheduling innovations have been necessary to accommodate the enlarged scale: twin parallel sessions run throughout the entire two-day event, and a poster format has been added (allowing more slots than "reserve papers").

The community deserves great credit for pitching in to meet this year's challenge of scaling up EMNLP within its traditional extremely tight submission-to-publication timeframe. To put this in perspective, the first pioneering meeting held in 1993 as the Workshop on Very Large Corpora (WVLC) featured 13 papers and a handful of reviewers. EMNLP was introduced in 1996 initially as a second annual event, then merged as EMNLP/VLC in 1999 and 2000 when empirical methodology penetrated the mainstream, after which the WVLC moniker was retired. In 2002 there were 53 reviewers. For EMNLP 2004 this year, there were 150+, nearly half of whom committed on extremely short notice. Moreover, members on average undertook about 50% more reviews than they signed up for. We were fortunate still to receive at least three reviews on the vast majority of papers (though technical obstacles prevented some from being entered into the conference workflow database).

The unprecedented level of submissions has also led us to introduce an Area Chair system—perhaps yet another step toward maturity for SIGDAT. Much credit goes to the Area Chairs for their dedication and citizenship in taking on a great deal of additional work with conscientious attention to detail, again with fairly little warning, in a year when the community had already borne unusually heavy reviewing duties from the large number of conferences.

We are happy to observe that the resulting program reflects a healthy continuing diversification of empirical NLP work. Increased representation is found in areas including semantic role labeling, discourse, dialog, summarization, question answering, sentiment, boundary detection, and sentence processing, in addition to traditional strengths in machine translation, parsing and tagging, word sense disambiguation, morphology, IR, and speech. We have striven to respect the reviewers as much as possible, with careful readings of their commentary together with the papers and not just the numeric ratings; we sought primarily to compensate as much as possible for reviewing variances and idiosyncrasies. But once again there were simply too many good papers submitted, and we can only hope you do enjoy the selection that the constrained program permitted!

This year we are introducing, as an experiment, a best paper award selection mechanism based on a larger (and hopefully, more statistically significant) sample size, so as to offset the larger number of papers. Rather than relying on the judgment of a small number of chairs, we are inviting all of you, the registrants, to vote for the paper you believe *will be seen ten years from now as having had the most impact.* A decade may be a long time to wait for confirmation, but aren't all good things worth waiting for...?

Still another program innovation this year is the joint keynote address with Senseval-3. As kindred spirits in empirical methodology, we trust that the coordination between our overlapping meetings will strengthen the experience for participants of both, and we are grateful to Rada Mihalcea and Phil Edmonds for their cooperative efforts. We are especially grateful to our speaker, Ken Church, for accepting this challenge.

Our thanks also go to David Yarowsky for assembling the proceedings, and to Toni Badia and

the local organizers for making possible the unexpected expansion of the conference.

The EMNLP special theme this year is Empiricism and Error Analysis, and we are happy to see a number of papers that made serious attempts to move forward on this front. We encouraged papers that apply existing or new methodology for meaningful empirical analysis of errors, independent of the application. The theme emphasizes that learning algorithms are not merely magic "black boxes". Corpus-based NLP has reached a stage of maturity where many competing models have reached comparable plateaus in performance. Merely reporting the aggregate accuracy or coverage percentages typically fails to uncover the model limitations that are fundamentally responsible for the plateaus. While reporting aggregate percentages remains essential, instructive research should also explain a model's limitations in more meaningful ways. This can be as simple as categorizing error statistics by finer-grained types of errors, to reveal specific areas of model limitations. Better yet, we seek deeper insight into the models' inherent representational biases, in the form of qualitative theoretical analyses that can be backed up by comparative quantitative measurements. We especially welcomed empirically verifiable implications arising for feature engineering or model selection, whether manual or automatic.

Welcome to EMNLP 2004—we hope you will find it thought-provoking and stimulating!

Dekang Lin, Google, lindek@google.com

Dekai Wu, HKUST, dekai@cs.ust.hk

Program Committee

Co-Chairs

Dekang Lin
Dekai Wu

Area Chairs

Daniel Gildea
Ido Dagan
Kenneth Church
Mary Harper
Pascale Fung
Yuji Matsumoto

Committee Members

Eneko Agirre
Srinivas Bangalore
Michele Banko
Regina Barzilay
Daniel Bikel
Jeff Bilmes
Chris Brew
Ted Briscoe
Bill Byrne
Sharon Caraballo
Claire Cardie
Marine Carpuat
John Carroll
Ciprian Chelba
John Chen
Colin Cherry
Stephen Clark
William Cohen
Michael Collins
Silviu Cucerzan
Walter Daelemans
Bonnie Dorr
Susan Dumais
Jason Eisner
Radu Florian
George Foster
Dayne Freitag
Jianfeng Gao
Ismael García Varea
Ralph Grishman
Jan Hajic
Marti Hearst
John Henderson
Graeme Hirst
Rebecca Hwa
Diana Inkpen
Franz Josef Och
Daniel Jurafsky
Frank Keller
Genichiro Kikui
Adam Kilgarriff
Dan Klein
Kevin Knight
Philipp Koehn
Greg Kondrak
Philippe Langlais
Guy Lapalme
Mirella Lapata
Gina Levow
Hang Li
Wei Li
Marc Light
Chin-Yew Lin
Paul Lu
Xiaoqiang Luo
Elliott Macklovitch
Rob Malouf
Gideon Mann
Daniel Marcu
James Mayfield
Andrew McCallum
Diana McCarthy
Dan Melamed
Rada Mihalcea
Dan Moldovan
Bob Moore
Lluís Màrquez
Masaaki Nagata
Hermann Ney
Grace Ngai
Kemal Oflazer
Miles Osborne
Lluís Padró
David Palmer
Patrick Pantel
Kishore Papineni
Ted Pedersen
Fernando Pereira
Fernando Pereira
Roberto Pieraccini
Massimo Poesio
Chris Quirk
Dragomir Radev
Owen Rambow
Giuseppe Riccardi
Stefan Riezler
German Rigau
Ellen Riloff
Eric Ringger
Roni Rosenfeld
Anoop Sarkar
Charles Schafer
Hinrich Schütze
Richard Sproat
Rohini Srihari
Suzanne Stevenson
Jian Su
Keh-Yih Su
Zhifang Sui
Simone Teufel
Christoph Tillmann
Erik Tjong Kim Sang
Nicola Ueffing
Takehito Utsuro
Antal van den Bosch
Stephan Vogel
Shaojun Wang
Andy Way
David Weir
Janyce Wiebe
Peng Xu
Kenji Yamada
Jun Zhao

Additional Reviewers

Laura Alonso i Alemany
Chris Brockett
Jan Bungeroth
Rich Campbell
Ann Copestake
Andres Corrada
Hal Daume III
Pablo Duboue
Michael Gamon
Roxana Girju
Sasa Hasan
Rose Hoberman
Theresa Jickels
Soo-Min Kim
Rob Koeling
Leah Larkey
Cong Li
Huifeng Li
Ping Li
Yang Liu
Wesley Mackay
David Martínez
Evgeny Matusov
Mark McLauchlan
Zhang Min
Daichi Mochihashi
Cheng Niu
Fuchun Peng
Deepak Ravichandran
Yihai Shen
Weifeng Su
David Vilar
Julie Weeds
Jerod Weinman
Theresa Wilson
Jerry Zhu

Table of Contents

PREFACE iii

TABLE OF CONTENTS vii

CONFERENCE PROGRAM xi

AUTHOR INDEX xv

Max-Margin Parsing
Ben Taskar, Dan Klein, Mike Collins, Daphne Koller and Christopher Manning 1

Active Learning and the Total Cost of Annotation
Jason Baldridge and Miles Osborne 9

Induction of Greedy Controllers for Deterministic Treebank Parsers
Tom Kalt 17

Unsupervised WSD based on Automatically Retrieved Examples: The Importance of Bias
Eneko Agirre and David Martinez 25

VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations
Timothy Chklovski and Patrick Pantel 33

Scaling Web-based Acquisition of Entailment Relations
Idan Szpektor, Hristo Tanev, Ido Dagan and Bonaventura Coppola 41

Bilingual Parsing with Factored Estimation: Using English to Parse Korean
David A. Smith and Noah A. Smith 49

Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM
Pascale Fung and Percy Cheung 57

Comparing and Combining Generative and Posterior Probability Models: Some Advances in Sentence Boundary Detection in Speech
Yang Liu, Andreas Stolcke, Elizabeth Shriberg and Mary Harper 64

Automatic Paragraph Identification: A Study across Languages and Domains
Caroline Sporleder and Mirella Lapata 72

Mixing Weak Learners in Semantic Parsing
Rodney D. Nielsen and Sameer Pradhan 80

Calibrating Features for Semantic Role Labeling
Nianwen Xue and Martha Palmer 88

Unsupervised Semantic Role Labelling
Robert S. Swier and Suzanne Stevenson 95

The Influence of Argument Structure on Semantic Role Assignment
Sebastian Padó and Gemma Boleda 103

Object-Extraction and Question-Parsing using CCG
Stephen Clark, Mark Steedman and James R. Curran 111

A Phrase-Based HMM Approach to Document/Abstract Alignment
Hal Daumé III and Daniel Marcu 119

Automatic Analysis of Plot for Story Rewriting
Harry Halpin, Johanna D. Moore and Judy Robertson 127

Mining Spoken Dialogue Corpora for System Evaluation and Modeling
Frederic Bechet, Giuseppe Riccardi and Dilek Hakkani-Tur 134

Monolingual Machine Translation for Paraphrase Generation
Chris Quirk, Chris Brockett and William Dolan 142

Verb Sense and Subcategorization: Using Joint Inference to Improve Performance on Complementary Tasks
Galen Andrew, Trond Grenager and Christopher Manning 150

Attribute-Based and Value-Based Clustering: An Evaluation
Abdulrahman Almuhareb and Massimo Poesio 158

The Leaf Path Projection View of Parse Trees: Exploring String Kernels for HPSG Parse Selection
Kristina Toutanova, Penka Markova and Christopher Manning 166

Incremental Feature Selection and l1 Regularization for Relaxed Maximum-Entropy Modeling
Stefan Riezler and Alexander Vasserman 174

A Distributional Analysis of a Lexicalized Statistical Parsing Model
Daniel M. Bikel 182

Adaptive Language and Translation Models for Interactive Machine Translation
Laurent Nepveu, Guy Lapalme, Philippe Langlais and George Foster 190

Improving Word Alignment Models using Structured Monolingual Corpora
Wei Wang and Ming Zhou 198

Phrase Pair Rescoring with Term Weighting for Statistical Machine Translation
Bing Zhao, Stephan Vogel, Matthias Eck and Alex Waibel 206

Dependencies vs. Constituents for Tree-Based Alignment
Daniel Gildea 214

A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources
Jiri Hana, Anna Feldman and Chris Brew 222

Applying Conditional Random Fields to Japanese Morphological Analysis
Taku Kudo, Kaoru Yamamoto and Yuji Matsumoto 230

A Hybrid Model for Morpho-Syntactic Annotation of German with a Large Tagset
Julia Trushkina and Erhard Hinrichs 238

Identifying Broken Plurals in Unvowelised Arabic Text
Abduelbaset Goweder, Massimo Poesio, Anne De Roeck and Jeff Reynolds.....246

NP Bracketing by Maximum Entropy Tagging and SVM Reranking
Hal Daumé III and Daniel Marcu.....254

Trained Named Entity Recognition using Distributional Clusters
Dayne Freitag.....262

Error Measures and Bayes Decision Rules Revisited with Applications to POS Tagging
Hermann Ney, Maja Popović and David Sündermann.....270

Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?
Hwee Tou Ng and Jin Kiat Low.....277

Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot
Ciprian Chelba and Alex Acero.....285

Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users
Silviu Cucerzan and Eric Brill.....293

A Boosting Algorithm for Classification of Semi-Structured Text
Taku Kudo and Yuji Matsumoto.....301

Learning to Classify Email into "Speech Acts"
William W. Cohen, Vitor R. Carvalho and Tom M. Mitchell.....309

The Entropy Rate Principle as a Predictor of Processing Effort: An Evaluation against Eye-tracking Data
Frank Keller.....317

Random Forests in Language Modeling
Peng Xu and Frederick Jelinek.....325

On Log-Likelihood-Ratios and the Significance of Rare Events
Robert C. Moore.....333

Learning Nonstructural Distance Metric by Minimum Cluster Distortions
Daichi Mochihashi, Genichiro Kikui and Kenji Kita.....341

From Machine Translation to Computer Assisted Translation using Finite-State Models
Jorge Civera, Elsa Cubel, Antonio L. Lagarda, David Picó, Jorge González, Enrique Vidal, Francisco Casacuberta, Juan M. Vilar and Sergio Barrachina.....349

Learning Hebrew Roots: Machine Learning with Linguistic Constraints
Ezra Daya, Dan Roth and Shuly Wintner.....357

LexPageRank: Prestige in Multi-Document Text Summarization
Güneş Erkan and Dragomir R. Radev.....365

A New Approach for English-Chinese Named Entity Alignment
Donghui Feng, Yajuan Lv and Ming Zhou.....372

Unsupervised Domain Relevance Estimation for Word Sense Disambiguation
Alfio Gliozzo, Bernardo Magnini and Carlo Strapparava 380

Statistical Significance Tests for Machine Translation Evaluation
Philipp Koehn 388

Instance-Based Question Answering: A Data-Driven Approach
Lucian Vlad Lita and Jaime Carbonell 396

TextRank: Bringing Order into Texts
Rada Mihalcea and Paul Tarau 404

Sentiment Analysis using Support Vector Machines with Diverse Information Sources
Tony Mullen and Nigel Collier 412

Evaluating Information Content by Factoid Analysis: Human annotation and stability
Simone Teufel and Hans van Halteren 419

Efficient Decoding for Statistical Machine Translation with a Fully Expanded WFST Model
Hajime Tsukada and Masaaki Nagata 427

Multi-Document Biography Summarization
Liang Zhou, Miruna Ticrea and Eduard Hovy 434