HLT-NAACL 2007

TextGraphs-2: Graph-Based Algorithms for Natural Language Processing

Proceedings of the Workshop

26 April, 2007
Rochester, NY, USA

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

TextGraphs-2 Workshop at HLT-NAACL 2007
was sponsored by the University of Leipzig, Germany

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

PREFACE

Recent years have shown an increased interest in bringing the field of graph theory into Natural Language Processing. In many NLP applications entities can be naturally represented as nodes in a graph and relations between them can be represented as edges. Recent research has shown that graph-based representations of linguistic units as diverse as words, sentences and documents give rise to novel and efficient solutions in a variety of NLP tasks, ranging from part of speech tagging, word sense disambiguation and parsing to information extraction, semantic role assignment, summarization and sentiment analysis.

This volume contains papers accepted for presentation at the TextGraphs-2 2007 Workshop on Graph-Based Algorithms for Natural Language Processing. This event took place on April 26, 2007, in Rochester, NY, USA, immediately following the HLT-NAACL Human Language Technologies Conference. It was the second workshop on this topic, building on the success of the first TextGraphs workshop at HLT-NAACL 2006. The workshop aimed at bringing together researchers working on problems related to the use of graph-based algorithms for Natural Language Processing and on the theory of graph-based methods. It addressed a broad spectrum of research areas to foster exchange of ideas and help to identify principles of using the graph notions that go beyond an ad-hoc usage. Unveiling these principles will give rise to applying generic graph methods to many new problems that can be encoded in this framework.

We issued calls for both regular and short, late-breaking papers. In total, ten regular and three short papers were accepted for presentation, considering the careful reviews of our program committee. We are indebted to all program committee members for their thoughtful, high quality and elaborate reviews, especially considering our extremely tight time frame for reviewing. The papers appearing in this volume have surely benefited from their expert feedback.

This year's workshop attracted papers employing graphs in a wide range of settings. While some contributions focus on analyzing the structure of graphs induced by language data or the interaction of processes on various levels, others use graphs as a means for data representation to solve NLP tasks, sometimes involving transformations on the graph structure.

H. F. Witschel introduces a new graph based meta model for Information Retrieval that subsumes many previous retrieval models and supports different forms of search. Improved unigram language models by a smoothing technique that accounts for word similarities are constructed by B. Jedynak and D. Karakos. Unsupervised grammar induction using latent semantics is the topic of A. M. Olney's research. V. Jijkoun and M. de Rijke view NLP tasks as graph transformations of labelled, directed graphs and experiment with tasks involving syntax and semantics. Syntactic dependency trees as a basis for semantic similarity are applied to textual entailment by D. Micol et al. D. Leite et al. find in their graph-based automatic summarization experiments that linguistic knowledge is necessary to improve automatic extracts. For multi-document summarization, evolving timestamped graphs are employed in the contribution of Z. Lin and M.-Y. Kan. A graph for the extraction of patterns combined with an extension of chance discovery is applied by C. S. Montero and K. Araki to human-computer dialogue mining. The small-world and scale-free property of linguistic graphs that go in hand with power-law distributions on entity and entity pair frequencies are examined in four papers: T. Zesch and I. Gurevych analyze the article and category graph of Wikipedia and measure correlation with WordNet. R. Ferrer

i Cancho et al. find correlations in the organization of syntactic dependency networks for a wide range of languages. Co-occurrence degree distributions are examined in a comparative study of Russian and English by V. Kapustin and A. Jamsen. In the setting of spell checking, M. Choudhury et al. find that spelling error probabilities for different languages are proportional to the average weighted degree of the corresponding SpellNet. A transductive classification algorithm based on graph clustering is described by K. Ganchev and F. Pereira, and tested on various NLP tasks.

Finally, having a prominent researcher as an invited speaker greatly contributes to the quality of the workshop. We thank Andrew McCallum for his talk and for the support that his prompt acceptance provided to the workshop.

Chris Biemann, Irina Matveeva, Rada Mihalcea and Dragomir Radev
April 2007

CHAIRS:

Chris Biemann, University of Leipzig, Germany
Irina Matveeva, University of Chicago, USA
Rada Mihalcea, University of North Texas, USA
Dragomir Radev, University of Michigan, USA

PROGRAM COMMITTEE:

Eneko Agirre, University of the Basque Country, Spain
Monojit Choudhury, Indian Institute of Technology
Diane Cook, Washington State University
Hal Daumé III, University of Utah
Gael Dias, Beira Interior University, Portugal
Güneş Erkan, University of Michigan
Michael Gamon, Microsoft Research
Bruno Gaume, IRIT, France
Andrew Goldberg, University of Wisconsin
Samer Hassan, University of North Texas
Hany Hassan, IBM, Egypt
Rosie Jones, Yahoo Research
Fabio Massimo Zanzotto, University of Rome, Italy
Andrew McCallum, University of Massachusetts Amherst
Ani Nenkova, Stanford University
Patrick Pantel, USC Information Sciences Institute
Uwe Quasthoff, University of Leipzig
Aitor Soroa, University of the Basque Country, Spain
Simone Teufel, Cambridge University, UK
Kristina Toutanova, Microsoft Research
Lucy Vanderwende, Microsoft Research
Dominic Widdows, Maya Design
Florian Wolf, F-W Consulting
Xiaojin Zhu, University of Wisconsin

INVITED SPEAKER:

Andrew McCallum, University of Massachusetts Amherst

WEBSITE:

http://www.textgraphs.org/ws07

Table of Contents

Conference Program

Thursday, April 26, 2007

8:45–9:00 Opening Remarks

Session 1: Session One

09:00–10:00 Invited Talk by Andrew McCallum

10:00–10:25 *Analysis of the Wikipedia Category Graph for NLP Applications*
Torsten Zesch and Iryna Gurevych

10:30–11:00 Coffee Break

Session 2: Session Two

11:00–11:25 *Multi-level Association Graphs - A New Graph-Based Model for Information Retrieval*
Hans Friedrich Witschel

11:25–11:50 *Extractive Automatic Summarization: Does more Linguistic Knowledge Make a Difference?*
Daniel S. Leite, Lucia H. M. Rino, Thiago A. S. Pardo and Maria das Graças V. Nunes

11:50–12:15 *Timestamped Graphs: Evolutionary Models of Text for Multi-Document Summarization*
Ziheng Lin and Min-Yen Kan

12:15–12:30 *Unigram Language Models using Diffusion Smoothing over Graphs*
Bruno Jedynak and Damianos Karakos

12:30–14:00 Lunch Break

Thursday, April 26, 2007 (continued)

Session 3: Session Three

14:00–14:25 *Transductive Structured Classification through Constrained Min-Cuts*
Kuzman Ganchev and Fernando Pereira

14:25–14:50 *Latent Semantic Grammar Induction: Context, Projectivity, and Prior Distributions*
Andrew M Olney

14:50–15:15 *Learning to Transform Linguistic Graphs*
Valentin Jijkoun and Maarten de Rijke

15:15–15:30 *Semi-supervised Algorithm for Human-Computer Dialogue Mining*
Calkin S. Montero and Kenji Araki

15:30–16:00 Coffee Break

Session 4: Session Four

16:00–16:25 *Correlations in the Organization of Large-Scale Syntactic Dependency Networks*
Ramon Ferrer i Cancho, Alexander Mehler, Olga Pustylnikov and Albert Diaz-Guilera

16:25–16:50 *DLSITE-2: Semantic Similarity Based on Syntactic Dependency Trees Applied to Textual Entailment*
Daniel Micol, Óscar Ferrández, Rafael Muñoz and Manuel Palomar

16:50–17:15 *How Difficult is it to Develop a Perfect Spell-checker? A Cross-Linguistic Analysis through Complex Network Approach*
Monojit Choudhury, Markose Thomas, Animesh Mukherjee, Anupam Basu and Niloy Ganguly

17:15–17:30 *Vertex Degree Distribution for the Graph of Word Co-Occurrences in Russian*
Victor Kapustin and Anna Jamsen