ACL-IJCNLP 2009

TextGraphs-4

2009 Workshop on Graph-based Methods for Natural Language Processing

Proceedings of the Workshop

7 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
`acl@aclweb.org`

ISBN 978-1-932432-54-1 / 1-932432-54-X

Introduction

The last few years have shown a steady increase in applying graph-theoretic models to computational linguistics. In many NLP applications, entities can be naturally represented as nodes in a graph and relations between them can be represented as edges. There have been extensive research showing that graph-based representations of linguistic units such 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 labeling, summarization, and sentiment analysis.

More recently, complex network theory, a popular modeling paradigm in statistical mechanics and physics of complex systems, was proven to be a promising tool in understanding the structure and dynamics of languages. Complex network based models have been applied to areas as diverse as language evolution, acquisition, historical linguistics, mining and analyzing the social networks of blogs and emails, link analysis and information retrieval, information extraction, and representation of the mental lexicon. In order to make this field of research more visible, this time the workshop incorporated a special theme on *Cognitive and Social Dynamics of Languages in the framework of Complex Networks*. Cognitive dynamics of languages include topics focused primarily on language acquisition, which can be extended to language change (historical linguistics) and language evolution as well. Since the latter phenomena are also governed by social factors, we can further classify them under social dynamics of languages. In addition, social dynamics of languages also include topics such as mining the social networks of blogs and emails. A collection of articles pertaining to this special theme will be compiled in a special issue of the *Computer Speech and Language* journal.

This volume contains papers accepted for presentation at the TextGraphs-4 2009 Workshop on Graph-Based Methods for Natural Language Processing. The event took place on August 7, 2009, in Suntec, Singapore, immediately following ACL/IJCNLP 2009, the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Being the fourth workshop on this topic, we were able to build on the success of the previous TextGraphs workshops, held as part of HLT-NAACL 2006, HLT-NAACL 2007 and Coling 2008. It aimed at bringing together researchers working on problems related to the use of graph-based algorithms for NLP and on pure graph-theoretic methods, as well as those applying complex networks for explaining language dynamics. Like last year, TextGraphs-4 has also been endorsed by SIGLEX.

We issued calls for both regular and short papers. Nine regular and three short papers were accepted for presentation, based on the careful reviews of our program committee. Our sincere thanks to all the 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 and we are therefore proud to present a very diverse program. We received quite a few papers on discovering semantic similarity through random walks. Daniel Ramage et al. explore random walk based methods to discover semantic similarity in texts, while Eric Yeh et al. attempt to discover semantic relatedness through random walks on the Wikipedia. Ameç Herdağdelen et al. describes a method for measuring semantic relatedness with vector space models and random walks.

Another set of papers were focused on popular graph-based machine learning techniques including classification and clustering. Swapna Somasundaran et al. employ opinion graphs for the purpose of polarity and discourse classification. Delip Rao and David Yarowsky propose a semi-supervised classification method on large scale graphs using map reduce. Yoshimi Suzuki and Fumiyo Fukumoto classify Japanese polysemous verbs using fuzzy C-means clustering method. Linlin Li and Caroline Sporleder discuss a cohesion graph based approach for unsupervised recognition of literal and nonliteral use of multiword expressions. Zheng Chen and Heng Ji propose a graph-based method for event coreference resolution. Scott Martens presents a quantitative analysis of treebanks using frequent subtree mining methods.

In the special theme category, we could select three papers. Sitabhra Sinha et al. present a paper pertaining to a topic that has recently been quite controversial. They show that a thorough network analysis reveals a structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions. Martijn Wieling and John Nerbonne discuss a bipartite spectral graph partitioning method to co-cluster varieties and sound correspondences in dialectology. David Ellis discusses social (distributed) language modeling, clustering and dialectometry.

Finally, we would like to thank Vittorio Loreto from University of Rome “La Sapienza” for his invited speech on “Collective dynamics of social annotation.” The talk was highly interesting and very pertinent to the special theme of the workshop. We are also grateful to Microsoft Research India for sponsoring the travel and accommodations for the invited speaker.

Monojit Choudhury, Samer Hassan, Animesh Mukherjee and Smaranda Muresan
August 2009

Organizers:

Monojit Choudhury, Microsoft Research (India)
Samer Hassan, University of North Texas (USA)
Animesh Mukherjee, Indian Institute of Technology (India)
Smaranda Muresan, Rutgers University (USA)

Program Committee:

Eneko Agirre, Basque Country University (Spain)
Edo Airoldi, Harvard University (USA)
Alain Barrat, C.N.R.S. (France)
Pushpak Bhattacharyya, IIT Bombay (India)
Chris Biemann, Powerset (USA)
Andras Csomai, Google Inc. (USA)
Hang Cui, Yahoo Inc (USA)
Hal Daume III, University of Utah (USA)
Mona Diab, Columbia University (USA)
Santo Fortunato, ISI Foundation (Italy)
Michael Gammon, Microsoft Research Redmond (USA)
Niloy Ganguly, IIT Kharagpur (India)
Lise Getoor, University of Maryland (USA)
Simon Kirby, University of Edinburgh (USA)
Ben Leong, University of Delaware (USA)
Vittorio Loreto, University of Rome ”La Sapienza” (Italy)
Irina Matveeva, Accenture Technology Labs (USA)
Alexander Mehler, Universitt Bielefeld (Germany)
Rada Mihalcea, University of North Texas (USA)
Roberto Navigli, University of Rome ”La Sapienza” (Italy)
John Nerbonne, University of Groningen (Netherlands)
Dragomir Radev, University of Michigan (USA)
Raghavendra Udupa, Microsoft Research (India)
Xiaojun Wan, Peking University (China)
Sren Wichmann, MPI for Evolutionary Anthropology (Germany)

Invited Speaker:

Vittorio Loreto, University of Rome ”La Sapienza” (Italy)

Table of Contents

Conference Program

Friday, August 7, 2009

	Session I: Opening
8:30–8:45	Inauguration by Chairs
8:45–9:48	Invited Talk by Prof. Vittorio Loreto
9:48–10:00	*Social (distributed) language modeling, clustering and dialectometry* David Ellis
10:00–10:30	Coffee Break
	Session II: Special Theme
10:30–10:55	*Network analysis reveals structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions* Sitabhra Sinha, Raj Kumar Pan, Nisha Yadav, Mayank Vahia and Iravatham Mahadevan
10:55–11:20	*Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology* Martijn Wieling and John Nerbonne
11:20–12:10	Panel Discussion
	Session III: Semantics
13:50–14:15	*Random Walks for Text Semantic Similarity* Daniel Ramage, Anna N. Rafferty and Christopher D. Manning
14:15–14:40	*Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering* Yoshimi Suzuki and Fumiyo Fukumoto
14:40–15:05	*WikiWalk: Random walks on Wikipedia for Semantic Relatedness* Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre and Aitor Soroa
15:05–15:18	*Measuring semantic relatedness with vector space models and random walks* Amaç Herdağdelen, Katrin Erk and Marco Baroni

Friday, August 7, 2009 (continued)

15:18–15:30 *Graph-based Event Coreference Resolution*
Zheng Chen and Heng Ji

15:30–16:00 Coffee Break

Session IV: Classification and Clustering

16:00–16:25 *Ranking and Semi-supervised Classification on Large Scale Graphs Using Map-Reduce*
Delip Rao and David Yarowsky

16:25–16:50 *Opinion Graphs for Polarity and Discourse Classification*
Swapna Somasundaran, Galileo Namata, Lise Getoor and Janyce Wiebe

16:50–17:15 *A Cohesion Graph Based Approach for Unsupervised Recognition of Literal and Non-literal Use of Multiword Expressions*
Linlin Li and Caroline Sporleder

17:15–17:40 *Quantitative analysis of treebanks using frequent subtree mining methods*
Scott Martens

17:40–18:00 Closing

Invited Talk
Collective Dynamics of Social Annotation

Vittorio Loreto
Dipartimento di Fisica, “Sapienza” Università di Roma,
Piazzale Aldo Moro 5, 00185 Roma, Italy
and
Complex Networks Lagrange Laboratory,
Institute for Scientific Interchange (ISI), Torino, Italy
vittorio.loreto@roma1.infn.it

The enormous increase of popularity and use of the WWW has led in the recent years to important changes in the ways people communicate. An interesting example of this fact is provided by the now very popular social annotation systems, through which users annotate resources (such as web pages or digital photographs) with text keywords dubbed tags. Collaborative tagging has been quickly gaining ground because of its ability to recruit the activity of web users into effectively organizing and sharing vast amounts of information. Understanding the rich emerging structures resulting from the uncoordinated actions of users calls for an interdisciplinary effort. In particular concepts borrowed from statistical physics, such as random walks, and the complex networks framework, can effectively contribute to the mathematical modeling of social annotation systems. First I will introduce a stochastic model of user behavior embodying two main aspects of collaborative tagging: (i) a frequency-bias mechanism related to the idea that users are exposed to each others tagging activity; (ii) a notion of memory, or aging of resources, in the form of a heavy-tailed access to the past state of the system. Remarkably, this simple modeling is able to account quantitatively for the observed experimental features with a surprisingly high accuracy. This points in the direction of a universal behavior of users who, despite the complexity of their own cognitive processes and the uncoordinated and selfish nature of their tagging activity, appear to follow simple activity patterns. Next I will show how the process of social annotation can be seen as a collective but uncoordinated exploration of an underlying semantic space, pictured as a graph, through a series of random walks. This modeling framework reproduces several aspects, so far unexplained, of social annotation, among which the peculiar growth of the size of the vocabulary used by the community and its complex network structure that represents an externalization of semantic structures grounded in cognition and typically hard to access.