COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA)

Workshop chairs:
Eva Hajičová, Lucie Poláková and Jiří Mírovský

15 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA)
Eva Hajičová, Lucie Poláková and Jiří Mírovský (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Preface

Discourse analysis as a domain focused on issues going beyond intra-sentential relations has been an important and intensively discussed topic of all prestigious linguistic and computational meetings, bringing important insights into the research area. Nevertheless we hope that our workshop on *Advances in Discourse Analysis and Its Computational Aspects* will not be just “one in the row” but will bring together and spread an up-to-date information on advanced computationally oriented studies in discourse analysis, and will provoke discussion on hot issues in the domain of the study in discourse, especially with respect to modern methodology and to computationally and corpus oriented research and its possible applications. Thus, we expect the workshop to attract a rather broad (and cross-domain) audience: those who are just starting their research in the given area will get enough stuff for thought how to proceed, and those who are in an advanced stage of their research will get a stimulating feedback from the floor and the discussion will make it possible for them to sharpen their ideas and plans.

In order to make these expectations real, the workshop program consists of two kinds of presentations: five invited position papers given by prominent researchers who have already had significant contributions to the field, and six papers selected during the anonymous review from those submitted by workshop participants. The topics of the position papers reflect the current state of the art and at the same time present a look ahead: *Aravind Joshi* (University of Pennsylvania, Philadelphia, USA), the founder and the head of the team which has offered the computational linguistic community one of the first comprehensive and most influential corpus of English annotated with discourse relations, the Penn Discourse TreeBank, opens a discussion on the specification of elements on which the annotation of discourse relations should rely, while *Nianwen (Bert) Xue* (Brandeis University, USA), who most unfortunately had at the last minute to cancel his personal attendance due to visa problems, in his abstract duly reminds us that a cross-lingual perspective introduces many not-yet or not-yet-fully explored phenomena that should be taken into account. *Kathleen McKeown* (Columbia University, New York, USA) documents that in language generation, discourse structure relations often play a prescriptive role in determining what to say next and she asks to which extent the annotation of the PDTB which couples discourse structure, syntactic structure and sense annotation offers an advantage over previous methods. *Kristiina Jokinen* (University of Helsinki, Finland and University of Tartu, Estonia) extends the discussion on information presentation to an interactive system with an important outreach to an application area. A non-negligible component part of the analysis and annotation of discourse relations is a cross-lingual computational study of anaphora accompanied by evaluation initiatives; the lessons learned during the experience with the annotation of the GNOME and ARRAU corpora of English, the LiveMemories corpus of Italian, and the ongoing annotation using the Phrase Detective game and the issues that still remain to be tackled – that is the subject of the position paper given by *Massimo Poesio* (University of Essex, Great Britain).

The timeline of the workshop program allows us to thematically group together the position papers with accepted presentations. We hope that such a grouping will help to concentrate on an intensive interaction and discussion of all the participants of the workshop. However, this does not mean that other relevant issues should be excluded from our discussion, both after the presentations and in the general discussion period at the end of the workshop.

Among the issues proposed to be discussed there are

- Intra-sentential and inter-sentential relations: commonalities and differences
- Explicit and implicit relations of coherence of discourse; means of implicit relations
- What can corpus annotation of discourse relations and related phenomena reveal?
- Annotation efforts undertaken in languages other than English, and their contribution to advances in Language Technologies and to a greater cross-linguistic understanding of coherence relations, their complexity and their lexicalization
- Advances in empirically-driven discourse-level methods of language processing (discourse parsing, sense detection) and their impact on theoretical understanding of discourse structure
- Discourse and dialogue, commonalities and differences (e.g. dialogue act standardization)
- Text segmentation and modeling of coherence in texts, tweets, dialogues, monologues etc.
- Structures other than coherence relations that discourse manifests (e.g. layout or "document structure"), or structure specific to particular genres (news report, scientific papers, errata, etc.)

We would like to thank all the invited speakers for their willingness to be with us at the workshop and to share their ideas with us, and also all the authors of the submissions for their contributions. We are most grateful to the Publication Chair Roger Evans for his most efficient efforts that helped us with the publication of the Workshop Proceedings and, last but not least, the local organizers of COLING 2012 for their continuous care of the COLING 2012 local organization for which they deserve a good measure of the credit.

Welcome to ADACA workshop at COLING 2012!

Eva Hajičová, Lucie Poláková, and Jiří Mírovský

ADACA organizers

Table of Contents

Abstracts of invited position papers

Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA)

Program

Saturday, 15 December 2012

10:00–10:20	Welcome, Introduction (Eva Hajičová)
10:20–11:00	*Remarks on some not so closed issues concerning discourse connectives (invited position paper)* Aravind Joshi
11:00–11:25	*Discourse Analysis of Sanskrit texts* Amba Kulkarni and Monali Das
11:25–12:00	Tea break
12:00–12:40	*Penn Discourse Treebank Relations and their Potential for Language Generation (invited position paper)* Kathleen McKeown
12:40–13:05	*Exploiting Discourse Relations between Sentences for Text Clustering* Nik Adilah Hanin Binti Zahri, Fumiyo Fukumoto and Suguru Matsuyoshi
13:05–13:30	*Measuring the Strength of Linguistic Cues for Discourse Relations* Fatemeh Torabi Asr and Vera Demberg
13:30–14:30	Lunch
14:30–15:10	*New Information in Wikitalk - story telling for information presentation (invited position paper)* Kristiina Jokinen
15:10–15:35	*Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT* Pavlína Jínová, Jiří Mírovský and Lucie Poláková
15:35–16:00	*Incremental Construction of Robust but Deep Semantic Representations for Use in Responsive Dialogue Systems* Andreas Peldszus and David Schlangen
16:00–16:30	Tea break
16:30–17:10	*Empirical methods in the study of anaphora: lessons learned, remaining problems (invited position paper)* Massimo Poesio
17:10–18:00	General discussion: Where we are and where to go