Studies in Computational Intelligence

Volume 677

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/7092

Leon R.A. Derczynski

Automatically Ordering Events and Times in Text

Springer

Leon R.A. Derczynski
Department of Computer Science
The University of Sheffield
Sheffield
UK

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-319-47240-9 ISBN 978-3-319-47241-6 (eBook)
DOI 10.1007/978-3-319-47241-6

Library of Congress Control Number: 2016953285

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Nanna Inie; heartfelt thanks for your attention, focus and passion.

Foreword

I am delighted to be able to write a few words of introduction to this new book on time and language. It is published at a very important time, in the midst of an explosion in artificial intelligence, where humans, hardware, data, and methods have combined at a fantastic rate to help not only us, but also our tools and computers, better understand our world.

Across the globe, in almost every language we encounter, we discover that we have evolved the ability to reason about time. Terms such as 'now' and 'tomorrow' describe regions of time; other terms reference events, such as 'opened' or 'hurricane'. This ability to refer to times or to events through language is important and gives humans much great ability in planning, storytelling, and describing the world around us. However, referring to events and times is not quite enough—we also need to be able to describe how these pieces all fit together, so that we can say when an event, like the 'hurricane', happened. This temporal structure can be thought of being built from relations that link each event and each time like a net. These temporal relations are encoded in the way we use language around events and times. Discovering how that code works, and what temporal relations a text is communicating to us, is the key to understanding temporal structure in texts.

Traditionally, computational linguistics—the study of computational techniques for language—has given the tools used to address automatic extraction of temporal information from language. Temporal information extraction typically involves identifying events, identifying times, and trying to link them all together, following patterns and relations in the text. One of the harder parts of this extraction process is linking together of events and times, to understand temporal structure. There have been many clever approaches to the task, from scholars and researchers in industry around the world. It is so hard that there has been, and still is, a long-running set of shared exercises, just for this: the TempEval challenges. The first of this series was proposed almost a decade ago in 2006 by me and my collaborators, which we started in order to advance temporal semantic annotation and the plethora of surrounding tasks.

Later, it was actually through one of these TempEval tasks that I first met Dr. Derczynski, and thereafter over many coffees and late dinners at venues like

LREC, or ISA, the semantic annotation workshop. Since, we have collaborated on temporal information extraction, co-organizing more recent TempEval tasks. Our current forthcoming work is a full-length textbook with Marc Verhagen on temporal information processing, with plenty of examples and thorough discussion of the multitude of issues in this fascinating and open area of science.

However, despite our and the community's years of work, and the heavy focus of many researchers through shared task series such as TempEval and i2b2, the problem of extracting temporal structure remains one of the hardest to solve in extracting temporal structure, and also the most important. Clearly, some fresh knowledge is needed.

This book adopts a different tactic to many others' research and describes a data-driven approach to addressing the temporal structure extraction problem. Based on a temporal relation extraction exercise involving systems submitted by researchers across the world, the easy and difficult parts of temporal structure are separated. To tell us where the hardest parts of the problem are, there is an analysis of the temporal relations that few or even none of the systems get right. Part of this analysis then attributes to various sources of linguistic information regarding temporal structure. Each source of information is drawn from a different part of linguistics or philosophy, incorporating ideas of, for example, Vendler, Reichenbach, Allen, and Comrie. The analysis then drives into the later parts of the book, where different sources of temporal structure information are examined in turn. Each chapter discussing a source of this information goes on to present methods for using it in automatic extraction, and bringing it to bear on the core problem: getting the structure of times and events in text.

My hope with this line of work is that it will bring some new knowledge about what is really going on with how temporal relations related to language. We can see the many types of qualitative linguistic theoretical knowledge compared with the hard reality of computational systems' outputs of temporal relations, and firm links emerge between the two. For example, we see links between iconicity—the textual order of elements in a document—and temporal ordering; or, an elegant validation of Reichenbach's philosophically based tense calculus, which, by including the progressive, ends up at Freksa's formal semi-interval logic almost by accident, while continuing to be supported by corpus evidence.

Bringing together all these threads of knowledge about time in language, while coupling them with empirically supported methods and evidence from the data that we have, has been a fruitful activity. This book advances work on some big outstanding problems, raising many interesting research questions along the way for both computer science and linguistics. Most importantly, it represents a valuable contribution to temporal information extraction, and thus to our overall goal: understanding how to process our human language.

June 2016

James Pustejovsky
TJX/Feldberg Chair of Computer Science
Department of Computer Science
Volen Center for Complex Systems
Brandeis University
Arlington, MA

Acknowledgements

A very special thanks to Robert Gaizauskas for his extensive help and guidance at many points; and to Yorick Wilks and Mark Steedman for their comments on an earlier version. The whole could not have been possible without the vision for the field and vast groundwork laid describing time in language, which is due to in great part to James Pustejovsky, as is gratitude for the foreword.

Finally, the book was produced during a period where I received support from, in no particular order: the EC FP7 project TrendMiner, the EC FP7 project Pheme, the EC H2020 project Comrades, an EPSRC Enhanced Doctoral Training Grant, the University of Sheffield Engineering Researcher Society, and the CHIST-ERA EPSRC project uComp.

Contents

List of Figures

List of Tables

Abstract

The ability to describe the order of events is crucial for effective communication. It is used to describe causality, to plan, and to relay stories. This temporal ordering can be expressed linguistically in a variety of ways. For example, one may use tense to describe the relation between the time of speaking and other events, or use a temporal conjunction to temporally situate an event relative to time. This ordering remains the hardest task in processing time in text. Very sophisticated approaches have yielded only small improvements over initial attempts. This book covers relevant background and discusses the problem, and goes on to conduct an analysis of temporal ordering information. This breaks the types of information used into different groups. Two major sources of information are identified that provide typing information for two segments: relations explicitly described by a signal word, and relations involving a shift of tense and aspect. Following this, the book investigates automatic temporal relation typing in both these segments, presenting results, introducing new methods, and generating a set of new language resources.