Studies in Computational Intelligence

Volume 589

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 world-wide distribution, which enable both wide and rapid dissemination of research output.

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

Roberto Basili · Cristina Bosco
Rodolfo Delmonte · Alessandro Moschitti
Maria Simi
Editors

Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project

Springer

Editors
Roberto Basili
Department of Computer Science, Systems and Production
University of Rome Tor Vergata
Rome
Italy

Cristina Bosco
Department of Computer Science
University of Turin
Turin
Italy

Rodolfo Delmonte
Department of Language and Cultural Studies, Department of Computer Science
Ca' Foscari University of Venice
Venezia
Italy

Alessandro Moschitti
Department of Computer Science and Information Engineering
University of Trento
Trento
Italy

Maria Simi
Department of Computer Science
University of Pisa
Pisa
Italy

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-319-14205-0 ISBN 978-3-319-14206-7 (eBook)
DOI 10.1007/978-3-319-14206-7

Library of Congress Control Number: 2014958283

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Portale per l'Accesso alle Risorse in Lingua Italiana (PARLI) is a project partially funded by the Ministero Italiano per l'Università e la Ricerca (PRIN 2008) from 2008 to 2012. The project was proposed by research units working in seven Italian universities, namely the University of Torino with a subunit at the University of Napoli "Federico II", the University of Pisa, the University of Roma "Tor Vergata", the University of Trento, the University of Venezia "Ca' Foscari". Moreover the Fondazione Bruno Kessler (FBK, Trento), the Istituto di Linguistica Computazionale "Antonio Zampolli"—CNR (Pisa) and the Associazione Italiana per l'Intelligenza Artificiale (AI*IA) played in the project the role of cooperating partners.

As the title of the project itself shows, PARLI mainly aimed at monitoring and fostering the harmonic growth and coordination of the activities of Italian NLP. In addition to that, it also proposes itself as a point of reference for the development of Italian NLP. According to this perspective, a web portal (http://parli.di.unito.it/) has been developed as a reference point for Italian NLP and for monitoring related activities. It includes links to existing resources and tools developed for Italian or applied to it. It mainly benefits from the data made available within the Evalita evaluation campaigns (http://www.evalita.it/) held in 2007, 2009 and 2011, and is linked by the NLP section of the AI*IA website (http://www.aixia.it/).

As far as the harmonic growth and coordination of Italian NLP is concerned, several activities promoted by PARLI members are attested by more than 50 publications, issued within the project, in international conferences, journals and workshops, among which are those related to the Evalita experiences, which were mainly organized and intensively participated by the PARLI members and cooperating partners.

There are several directions in which research on NLP has made considerable progress in the last few years. The chapters collected in this volume are selected as a sample of those performed for Italian NLP and especially oriented to the goals of the PARLI project, namely the consolidation and harmonization of existing linguistic resources, the development of new resources and tools that can harmonically operate and grow together, and the study of models for the comparison and evaluation of tools and resources.

Even if more and more treebanks are currently available also for lesser studied languages, none of the existing resources for Italian is large enough to train and test NLP systems with high reliability. This is also because they are featured by annotations which are far from standards applied in larger and well-known data sets. The consolidation and harmonization of these existing linguistic resources is at issue in the article of Simi, Montemagni and Bosco, where a methodology for merging and converting treebanks in a standard annotation format is designed. The format is applied to two existing Italian resources, i.e. Turin University Treebank (TUT) and ISST–TANL, in order to build a larger data set in the standard de facto Stanford Dependency format. Also, the contribution of Delmonte refers to issues related to standards for annotation. It highlights a peculiar limit of the formats of resources on which state-of-the-art parsers are currently trained, i.e. the exclusion of null elements, and faces the problems derived from the conversion in a format almost semantically complete which includes null elements.

The development of new resources that can grow and cooperate together is the topic of the contribution of Sanguinetti, Lesmo and Bosco, where a recently released parallel treebank is proposed for cross-linguistic comparisons among Italian, English and French, and a study for the development of a dependency-based alignment system. This resource applies the same format of the TUT and takes advantage of the tools developed for this treebank; in addition, it can influence machine translation as well as linguistic investigations. Another kind of approach is taken in the chapter by Magnini, Zanoli and Firoj, which presents a comparative analysis of named entities extraction from both written and spoken documents, thus introducing a new perspective related to spoken language.

The contributions of Croce, Basili and Moschitti and the that by Croce, Filice and Basili describe the development of tools and related methodologies. The former chapter tackles the definition and evaluation of the semantically Smoothed Partial Tree Kernel, which is a generalized formulation of one of the most performant Convolution Kernels, i.e. the Tree Kernel, by extending the similarity between tree structures with node similarities. The latter chapter instead discusses a perspective centred on Convolution Kernels and the formulation of a Partial Tree Kernel that integrates syntactic information and lexical generalization, in order to define methods able to express the meaning of phrases or sentences as operations on lexical representations.

The contribution of Alicante, Bosco, Corazza and Lavelli and the that by Mazzei deal with the study of models for comparison and evaluation of tools and resources. The former chapter is a collection of parsing experiments performed on TUT data in order to compare the two main paradigms, i.e. dependency and constituency, and forms of annotation featured by a different amount of linguistic knowledge. In the chapter by Mazzei, instead, an ensemble system for dependency parsing of Italian is presented where three parsers known in the literature are separately trained and combined by means of a majority vote on a common data set.

According to the spirit of the project PARLI, the resources and tools created within the project or made available by their partners are freely distributed. Moreover, as attested also by the richness of the future directions drawn in the

chapters here collected, it should be desirable that the activities associated with PARLI do not terminate at the end of the funded project itself. PARLI, the portal and the resources associated with it should continue to be managed even later, hoping they could be a key factor in resource development in computational linguistics for Italian and beyond.

Roberto Basili
Cristina Bosco
Rodolfo Delmonte
Alessandro Moschitti
Maria Simi

Acknowledgment

We all thank and remember Leonardo Lesmo, coordinator of the PARLI project and valuable teacher, colleague and friend.

Contents

Contributors

Firoj Alam SIS Lab, Department of Information Engineering and Computer Science, University of Trento, Povo (TN), Italy

Anita Alicante Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Università di Napoli Federico II, Naples, Italy

Roberto Basili Department of Computer Science, Systems and Production, University of Roma Tor Vergata, Rome, Italy

Cristina Bosco Dipartimento di Informatica, Università di Torino, Torino, Italy

Anna Corazza Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Università di Napoli Federico II, Naples, Italy

Danilo Croce Department of Computer Science, Systems and Production, University of Roma Tor Vergata, Rome, Italy

Rodolfo Delmonte Department of Language and Cultural Studies, Department of Computer Science, Ca' Foscari University of Venice, Venezia, Italy

Simone Filice Department of Computer Science, Systems and Production, University of Roma Tor Vergata, Rome, Italy

Alberto Lavelli HLT Research Unit, Fondazione Bruno Kessler, Povo, TN, Italy

Bernardo Magnini FBK-irst, Povo (TN), Italy

Alessandro Mazzei Dipartimento di Informatica, Università di Torino, Torino, Italy

Simonetta Montemagni Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR), Pisa, Italy

Alessandro Moschitti Department of Computer Science and Information Engineering, University of Trento, Povo (TN), Italy

Manuela Sanguinetti Dipartimento di Informatica, Università di Torino, Torino, Italy

Maria Simi Dipartimento di Informatica, Università di Pisa, Pisa, Italy

Roberto Zanoli FBK-irst, Povo (TN), Italy